Defense Threat Reduction Agency
8725 John J. Kingman Road, MS
6201 Fort Belvoir, VA 22060-6201

DTRA-TR-16-72

TECHNICAL REPORT

Understanding the fundamental principles underlying the survival and efficient recovery of multi-scale Techno-Social Networks following a WMD event (A)

Distribution Statement A. Approved for public release, distribution is unlimited.

July 2016

HDTRA1-09-1-0039

Zoltan Toroczkai

Prepared by:
University of Notre Dame Du Lac
940 Grace Hall
Notre Dame, IN 46556

DESTRUCTION NOTICE:

Destroy this report when it is no longer needed.
Do not return to sender.

PLEASE NOTIFY THE DEFENSE THREAT REDUCTION AGENCY, ATTN: DTRIAC/ J9STT, 8725 JOHN J. KINGMAN ROAD, MS-6201, FT BELVOIR, VA 22060-6201, IF YOUR ADDRESS IS INCORRECT, IF YOU WISH IT DELETED FROM THE DISTRIBUTION LIST, OR IF THE ADDRESSEE IS NO LONGER EMPLOYED BY YOUR ORGANIZATION.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*
00-07-2016

2. REPORT TYPE
Final Report

3. DATES COVERED *(From - To)*
05.01.2009 -- 08.30.2015

4. TITLE AND SUBTITLE
Understanding the fundamental principles underlying the survival and efficient recovery of multi-scale Techno-Social Networks following a WMD event (A)

5a. CONTRACT NUMBER
HDTRA1-09-1-0039

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Prof. Zoltan Toroczkai, Interdisciplinary Center for Network and Departments of Physics, Computer Science and Engineering, University of Notre Dame

Prof. Alessandro Vespignani, Laboratory for the Modeling of Biological and Socio-technical Systems, College of Computer Sciences, Northeastern University

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Notre Dame Du Lac,
Office of Research
940 Grace Hall, Notre Dame, IN 46556-5602

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Defense Threat Reduction Agency
8725 John J. Kingman Road
Fort Belvoir, VA 22060-6201

10. SPONSOR/MONITOR'S ACRONYM(S)
DTRA

11. SPONSOR/MONITOR'S REPORT NUMBER(S) DTRA-TR-16-72

12. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution Statement A. Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Understanding the fundamental principles underlying the robustness of Techno-Social Networks (TSN-s) is a tall, but inevitable challenge. As the society becomes increasingly dependent on infrastructure networks, the implications of large-scale events (such as WMD), become increasingly important as illustrated by the recent Ebola and Zika outbreaks. Such events serve as warnings about the vulnerabilities lying in strongly interconnected systems. The performed work addressed directly several of the announced call's objectives, by developing an understanding for the fundamentals of networks in order to help develop methods that optimize the robustness of these systems.

15. SUBJECT TERMS
Complex techno-social networks, data driven computational modeling, system behavior prediction, vulnerability detection

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Paul Tandy
					19b. TELEPHONE NUMBER *(Include area code)*
UU	UU	UU	SAR	122	703-767-4663

UNIT CONVERSION TABLE

U.S. customary units to and from international units of measurement[*]

U.S. Customary Units	Multiply by → ← Divide by[†]	International Units
Length/Area/Volume		
inch (in)	2.54 × 10^{-2}	meter (m)
foot (ft)	3.048 × 10^{-1}	meter (m)
yard (yd)	9.144 × 10^{-1}	meter (m)
mile (mi, international)	1.609 344 × 10^{3}	meter (m)
mile (nmi, nautical, U.S.)	1.852 × 10^{3}	meter (m)
barn (b)	1 × 10^{-28}	square meter (m^2)
gallon (gal, U.S. liquid)	3.785 412 × 10^{-3}	cubic meter (m^3)
cubic foot (ft^3)	2.831 685 × 10^{-2}	cubic meter (m^3)
Mass/Density		
pound (lb)	4.535 924 × 10^{-1}	kilogram (kg)
unified atomic mass unit (amu)	1.660 539 × 10^{-27}	kilogram (kg)
pound-mass per cubic foot (lb ft^{-3})	1.601 846 × 10^{1}	kilogram per cubic meter (kg m^{-3})
pound-force (lbf avoirdupois)	4.448 222	newton (N)
Energy/Work/Power		
electron volt (eV)	1.602 177 × 10^{-19}	joule (J)
erg	1 × 10^{-7}	joule (J)
kiloton (kt) (TNT equivalent)	4.184 × 10^{12}	joule (J)
British thermal unit (Btu) (thermochemical)	1.054 350 × 10^{3}	joule (J)
foot-pound-force (ft lbf)	1.355 818	joule (J)
calorie (cal) (thermochemical)	4.184	joule (J)
Pressure		
atmosphere (atm)	1.013 250 × 10^{5}	pascal (Pa)
pound force per square inch (psi)	6.984 757 × 10^{3}	pascal (Pa)
Temperature		
degree Fahrenheit (^{o}F)	[T(^{o}F) − 32]/1.8	degree Celsius (^{o}C)
degree Fahrenheit (^{o}F)	[T(^{o}F) + 459.67]/1.8	kelvin (K)
Radiation		
curie (Ci) [activity of radionuclides]	3.7 × 10^{10}	per second (s^{-1}) [becquerel (Bq)]
roentgen (R) [air exposure]	2.579 760 × 10^{-4}	coulomb per kilogram (C kg^{-1})
rad [absorbed dose]	1 × 10^{-2}	joule per kilogram (J kg^{-1}) [gray (Gy)]
rem [equivalent and effective dose]	1 × 10^{-2}	joule per kilogram (J kg^{-1}) [sievert (Sv)]

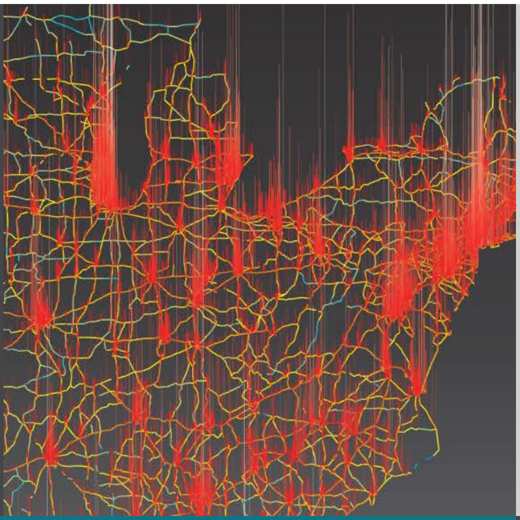
[*] Specific details regarding the implementation of SI units may be viewed at http://www.bipm.org/en/si/.

[†] Multiply the U.S. customary unit by the factor to get the international unit. Divide the international unit by the factor to get the U.S. customary unit.

2015-11-16

HDTRA1-09-1-0039

Final Report



Understanding the fundamental principles underlying the survival and efficient recovery of multi-scale Techno-Social Networks following a WMD event (A)

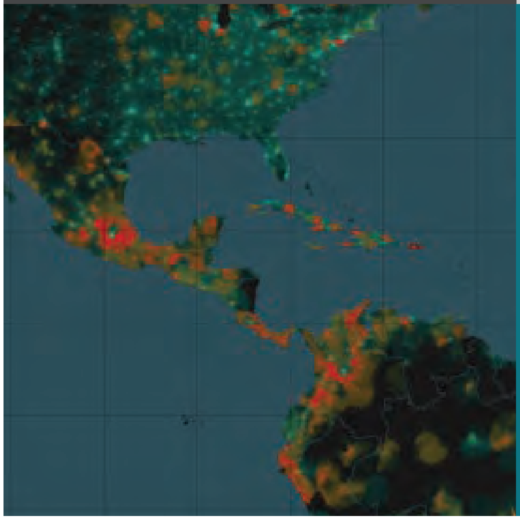
Principal Investigators

Zoltan Toroczkai

Interdisciplinary Center for Network Science and Applications (iCeNSA) & Departments of Physics, Computer Science and Engineering, University of Notre Dame

Alessandro Vespignani

Laboratory for the Modeling of Biological and Socio-technical Systems, College of Computer Sciences, Northeastern University



This report covers the period
05.01.2009 – 08.30.2015

Contents

I. Executive summary

Motivation: Understanding the fundamental principles underlying the functional robustness of Techno-Social Networks (TSN-s) is a tall, but inevitable challenge. As the society becomes increasingly dependent on infrastructure transport networks, the implications of large-scale disruptions (e.g., due to WMD events) of these systems also become increasingly important. In our ever more connected and interdependent world, small perturbations can have far reaching effects. Large-scale events, such as cascading failures (e.g., the East coast blackout on Aug. 14, 2003) or epidemics (SARS and most recently Ebola) are typical to complex networks, and they serve as warnings about the type of behaviors one can expect in strongly interconnected systems. As a result, we demand ever-increasing predictive power and ability to evaluate the impact of potential failures. There has been, however, little to no mathematical and analytical methodology that would allow a realistic assessment of the consequences of large-scale failures in critical infrastructure TSN-s, largely due to two major aspects of these networks: *i) Multivariate and heterogeneous constraints.* Real-world TSNs typically operate under a wide range of constraints and limitations imposed both from the technical (physical) side and the social (usage) side. Physical constraints include geographical embeddedness; energy limitations (transportation costs, battery life); device capacity, latency and bandwidth, while usage constraints include software limitations, communication protocols and regulations. To advance empirical applications, network models will need to address these constraints since they strongly influence the response and failure modes of a network in case of WMDs. The dynamics of constrained complex networks, especially if they are directed, weighted and hierarchical are not well understood. The adaptive capabilities, including recovery of TSNs facing major destructive events has been a relatively uncharted territory.

ii) Multiscale nature. Most of the TSNs are characterized by inherently complex processes involving many different scales in time and space, as well as in the operational and dynamical attributes of network elements. These features represent major challenges in our modeling and predictive understanding of TSNs. When is it possible to average small scales into large scale effects? When the addition of details and new scales induce a change in the survival and recovery behavior of the network? How do the various scales suffer and adapt to the same damage?

Addressing some of these challenges, our project has developed:

A. *Quantitative multiscale methods* for network analysis using large-scale real-world TSN datasets;
B. Efficient algorithms for TSN *vulnerability and robustness analysis*;
C. Methods for *subsystem diagnosis* with limited or no information available directly from the subsystem;
D. Studies into efficient, distributed and adaptive defense and recovery strategies for massively damaged or compromised TSNs, using simulations on synthetic networks.

Background: Today's world is supported by organically evolving network structures, which, transport and store various entities, including materials, energy, information and people across vastly varying spatial and temporal scales. Although physical transport networks have been around for a while, communication and computing infrastructure networks have been forming a dense web around those in the past two decades. All these networks are used and driven by humans, and thus by the network of relations among them, the Social Network (SN). While the technological advancements provide us with tools to better measure and quantify social interactions, they, on the other hand influence social networks by providing novel ways for human interactions. Technology and social organization mutually shape each other leading to the formation of multiscale dynamic techno-social networks (TSNs). This human-machine symbiosis and mutual exploitation becomes even more accentuated in network centric warfare and battlefield systems. The role of networks in warfare moves increasingly from a supporting position to becoming the weapon itself. As threats relocate from open battlefields towards more covert urban areas, the theater of operations becomes very heterogeneous, requiring global surveillance all the time. Engaging enemies under these circumstances is only possible by networked systems that monitor, record, assess and adapt to situations in real-time by integrating information from all scales, as envisioned by the Global Information Grid (GIG) project of DoD. Such networks necessarily have to possess flexible connectivity, ideally to be able to self-optimize, self-diagnose and self-heal. While the studies performed in this work with data coming from the civilian sector of networks, the fundamental conclusions, along with the mathematical results, methods and algorithms developed here are applicable to DoD networks as well.

Real-world Techno-Social Networks (TSN): These networks are heterogeneous, multiscale, with both static and mobile components and can be characterized by 7 major elements and the interactions among them:

1. *Spatially distributed Resource-Locations (RLs)*. These are spatial regions where people spend most of their time, including buildings in a city, cities within a county, counties within states, and finally countries, *spanning scales* from meters to thousands of km-s. The distribution of RLs is heterogeneous, fractal-like, determined by the distribution of resources, natural and man-made.
2. *Transportation networks (TN)*. These connect the RLs, and thus span multiple scales: from streets, to roadways, to highways, and finally to airline routes.
3. *Humans as transport agents*. People are in constant motion between the RLs, using the transportation routes and represent the flow in the transportation network.
4. *Human proximity networks (PN)*. Most of the time, people are co-located with other people (within the RLs), forming dynamic proximity networks, which serve as substrate to the spread of bio-pathogens.

5. *Social networks (SN)*. This is the network of "who knows who" and it drives/influences the communication as well as the physical mobility of people on the transportation network.
6. *Communication Infrastructure Networks (CIN)*. A large portion of information is transmitted via fixed infrastructures including the Internet, wire-line phone/base-station and the satellite network. The user nodes are at RLs, with links forming a complex, multiscale network, *different from TN*.
7. *Mobile Communication Networks (MCN)*. Mobile devices allow en-route communications between people. Since cell-phone usage is logged and cell-phones can be geo-located, they serve as massive (in volume), longitudinal *probes both into the social network and the mobility patterns of people*. Ad-hoc mobile networks use the (ad-hoc) connectivity between the radio ranges of devices with broadcast capabilities to transmit messages.

WMD event classification from a network perspective: In the following we consider two major types of WMD events: a) Type *G-WMD* which corresponds to massive, but *geo-localized* events such as nuclear, radiological or massive release of bio or chemical agents (anthrax, mustard, Sarin, etc.). and b) Type *E-WMD*, which is a bio-agent (e.g., smallpox) covertly introduced into the population to induce wide-spread *epidemic*; These events affect TSNs in very different but specific ways, and they must be treated accordingly. However, the tools developed in this project will allow the analysts to test in-silico for arbitrary failure scenarios.

TSN-s and their defense present the ultimate challenge in network research due to the number of coupled components, their relationships and the wide range of temporal and spatial scales involved. The physical and social components are intertwined and they influence each other especially under WMD stressors. The consequences of major events are not localized to the attack region only: information spreads almost instantaneously on the network modifying both the mobility and communication patterns of people, resulting in network overload in the transportation (TN) and as well as the communication (CIN/MCN) components, and potentially their failure.

Relevance: The performed work addresses directly several of the announced objectives of the call under Topic A, namely "Advancing the knowledge of network theory for understanding robustness" by developing an understanding for the fundamentals of physical and social networks in order to help develop methods that optimize the robustness of these systems and to provide tools that predict the consequences of WMD caused disruptions. The tasks presented here methodically tackle the research areas of interest of the call, as follows: The development of efficient methods and tools for topological and dynamical analysis of composite physical/social networks; The development of synthetic models which faithfully capture the essential structural and dynamical features of multiscale composite networks including their vulnerability and robustness properties; Mathematical formulation of these features into laws that govern the dynamics of these systems across all the scales; Validation of the results on several real-world large-scale TSN networks; Development of computational tools for assessing TSN response to various damage scenarios, including E(G)-WMD events. The real-world network data used for research in this project are actually part of the existing human infrastructures for communication and transport. Thus our results on robustness, vulnerability and adaptive recovery, and in case of a WMD event, our recommendations and guidance should apply directly to these real-world systems. Next we briefly highlight some of the work done during this project, that has direct relevance to DoD, from both the G-WMD and E-WMD areas, in particular, on massive transportation networks and contagion spread such as Ebola. However, these highlighted results were enabled by a large body of technical results that can be found in the rest of the document and in the publications generated. These technical results are both mathematical and algorithmic in nature and their applicability extends beyond DoD applications, they are valuable for the whole of network science community. As network science is arguably the most interdisciplinary subject (as many, wildly different systems can be represented by networks), our results are applicable to a wide range of subject areas and topics (some of our methods e.g., have been applied in neuroscience).

G-WMD related research highlights

Spatial transportation networks, and in particular, the highway transportation network forms the vascular system of our society. It has evolved organically, and its properties encode the modalities in which us and our economy interacts across space, ensuring the vitality of our society. Understanding the flow dynamics in such networks is crucial for many reasons, including the prediction of epidemic patterns (human mobility patterns determine the evolution of disease spread) and the smart design of spatial infrastructure networks. From a DoD perspective, such an understanding is key for assessing the effects of WMD induced disruptions to the transportation system, and on the flipside, helping with generating strategies of attack on adversarial transportation networks. As we show in our work, the effects of a geo-localized WMD attack (such as a

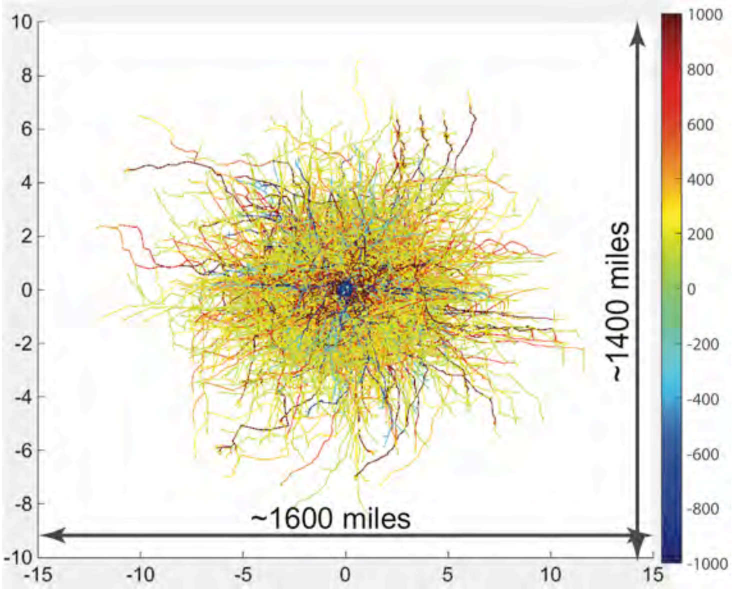


Fig I. Changes in average daily traffic in the highway transportation network, due to a G-WMD event (center of image) approximately equivalent to the detonation of a B53 strategic nuclear weapon in a single location. The image overlays the effects from 1000 such G-WMD events with epicenters chosen at random locations. The figure shows that there are specific locations in the network, where such an event would have significant and far reaching (spanning to half the US) consequences.

nuclear event) are not constrained only to within the immediate vicinity of the network, but it can spread network-wide, much like cracks in a brittle material can propagate throughout the whole material, in spite the fact that their origin may be a point-like shock. While such long-range propagation of effects is unsurprising in networks that possess many shortcuts, such as in airline transportation (we are all familiar with the fact that delays in a major airport can have ripple effects on the whole network), the fact that they also occur in roadway transportation (which is is where massive transportation of goods, essential for the functioning of a whole society, takes place) is rather surprising and counter-intuitive, because roadway networks are spatial networks without shortcuts. Figure I. shows the changes in average daily traffic after the detonation of a B53 strategic nuclear weapon of 9MT yield, rendering the transportation network unusable and unapproachable within a 20km radius region around the epicenter. In this plot we overlaid the effects from 1000 such in-silico experiments corresponding to 1000 epicenters chosen approximately at random from the continental US. As one can see, significant effects from some G-WMD events can be registered even out to 600 miles from the epicenter, effectively spanning half the US. The effects of such G-WMD events need to be studied both in volume and extent. While, clearly, high density population areas, such as metropolitan areas present vulnerabilities en-mass, we have found that, contrary to intuition, there are hot-spots in the network, where such events would have significant long-range effects in spite the fact the epicenters themselves may have little or no population. If these long-range effects overlap with roadway segments used for transportation of critical materials (e.g., they have to have a certain width), then as a response the system has to invest in providing alternative, adequate routes. However, building new roadways, is not only costly, but also takes a certain amount of time, and these can be significantly disruptive effects. A crucial task is then finding the circumstances or properties under which an epicenter generates long-range effects. To be able to provide answers to such questions, one needs to have a first-principles based understanding of the roadway transportation network, one that is not based on fitting parameters. The reason for this is that once the network changes (due to G-WMD events), the fitting parameters are also expected to change, and we cannot know those *a-priori*. Moreover, for this knowledge to be useful, it cannot be based on fitting parameters characteristic to one network or one country. Thus, before we tested disruption scenarios, we needed first to understand the flow dynamics in roadway transportation networks, as determined by the distribution of the population, the existing network of highways and human mobility properties.

The quantitative prediction of flows requires the solution of two fundamental problems. One is of mainly socio-demographic nature and it entails determining the number of individuals travelling between two given locations (The Mobility Law) and the other is a network distribution problem (The Flux Distribution problem) and it entails assigning the network paths to the individuals that are travelling between the locations. The first problem traditionally has been treated using the so-called gravity law and most recently, the *radiation law*. The radiation law [1] was derived from

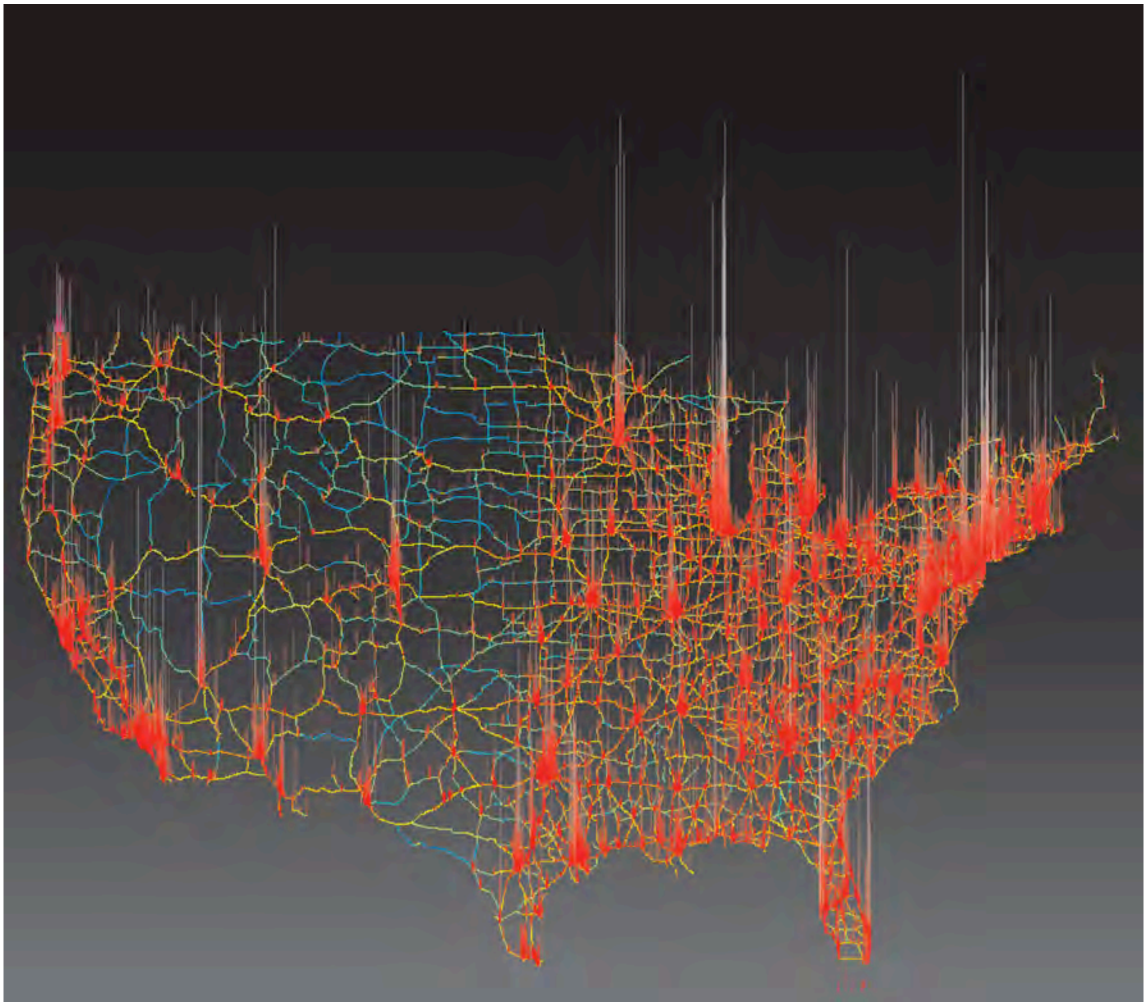


Fig II. Coupling between population distribution and roadway network flows in the continental US.

a simple social-demographic model of job-search and it has been shown to describe much better the number of people wanting to travel between two sites than the gravity law, and unlike the gravity law (and its variants) it has no fitting parameters. However, as we have shown, it is inadequate to predict flows in networks, as the radiation model does not couple the flow dynamics with the underlying network. The Flux Distribution problem is a network problem, but it has a social aspect as well, in that it depends on the travel cost criterion that individuals use when selecting a travel path between source and destination. There are two main criteria in this case, one being travel distance and the other being travel time. In our work we show that these two fundamental problems cannot be treated separately if one wants to realistically predict network flows, see Fig II. We demonstrate that the Mobility Law must take into account the cost of travel *on the network*. Our model includes a general cost function that allows us to compare predictions based on arbitrary cost criteria for travel, including those based on travel distance and travel time. Using a large-scale US highway network dataset from MIT-s database for validation, we show that the travel time based cost predictions achieve a *significantly better* agreement with real traffic than those based on travel distance alone. This indicates that people preponderantly choose paths based on time to destination rather than distance to destination. While this is intuitive, we are demonstrating it quantitatively for the first time. Applying our model on the US roadway network database we discovered that the mobility fluxes obey a self-similar scaling law that holds over seven orders of magnitude. We show that this scaling form is universal and derive its exponent (1.5). We show both numerically and analytically that when using the original radiation law (that is when the network structure is not taken into account), the corresponding scaling exponent becomes 1.3 rather than the correct 1.5. We have then exploited this scaling law to formulate an efficient flux distribution algorithm (the 2nd problem). We have published these results in Nature Communications [2].

From a mathematical and algorithmic point of view we have contributed with several results, as described in the rest of the report. These include methods for modeling networks based on partial information on their structure, such as degree-based network construction and sampling and maximum entropy based modeling methods. Related to the latter we have solved the degeneracy problem that was open for nearly 30 years. Through our range-limited betweenness algorithm we have provided a multiscale decomposition of the distribution of paths in a network. Additionally, we have generated the mathematics and accompanying efficient algorithms that can identify the structural vulnerability backbone of a given network. This backbone is the smallest structure, which when damaged, caused the largest disruption in the network (the shattering problem). The latter algorithms have seen translational applications, for example, in brain neuronal networks.

E-WMD related research highlights

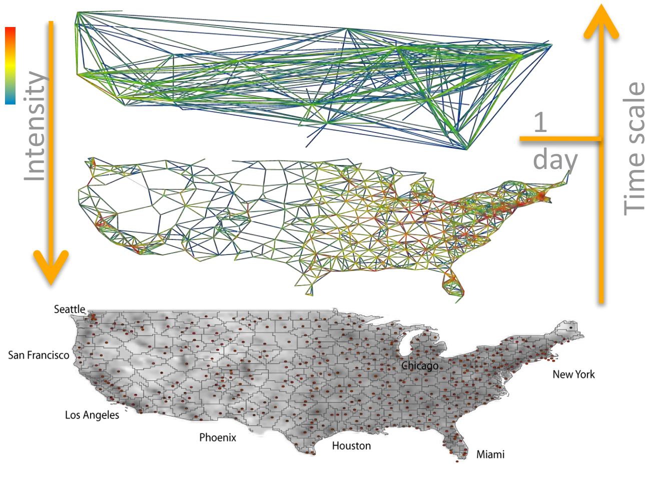


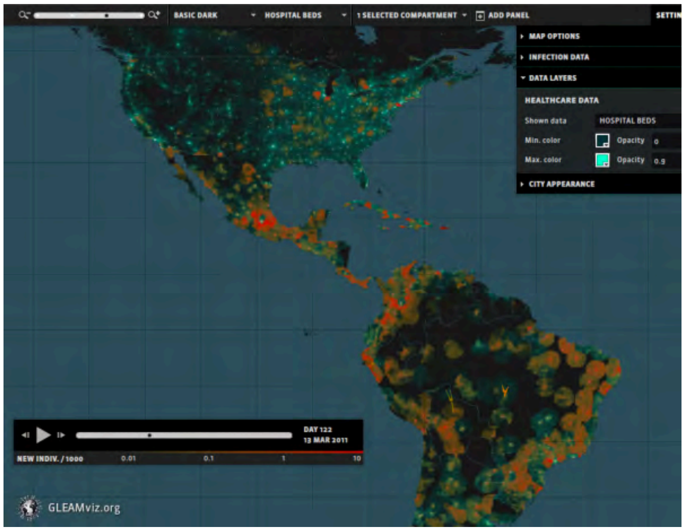
Fig. III. : Multiscale mobility and population networks.

In our project we have focused on a number of issues related to bio-agent or Epidemic (e.g., smallpox) events - E-WMD - covertly introduced into the population to induce wide-spread epidemic. The work was focused on providing non-incremental advances to our understanding and predictability power for the resilience and robustness of complex interconnected techno-social networks in the case of threats such as the spreading of highly pathogenic agents. This implies coping with the inherent multiscale nature of techno-social networks, and face the challenge of developing novel basic mathematical approaches to theories and algorithms able to cope with a wide range of scale-mixing dynamics; from the local commuting of people to international travel, from the individual to the population level dynamic of biological processes.

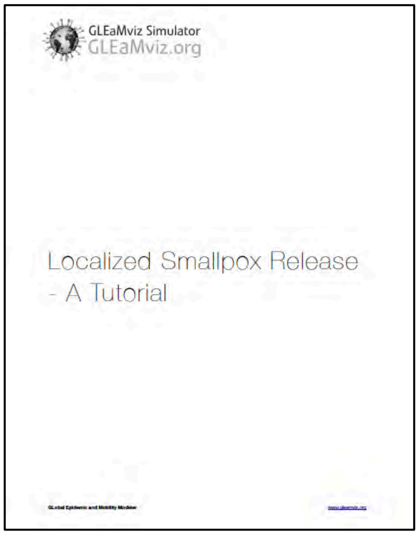
In the case of E-WMD events we have tackled the modeling challenges by developing a general analytical framework for the understanding of *multiscale interactions in reaction-diffusion models*. In particular, we developed a method for the multiscale integration of mobility networks in the analysis of potentially pandemic pathogens spread. The results published in several papers, including PNAS [3], and Science [4], defined the basic structure of a general computational platform – the global Epidemic and mobility model (GLEAM) - for the study of global pandemic and E-WMD events.

In order to address the *resilience of structured population to EMWD* we have tackled a number of specific features that characterizes how contagion processes transition from local (contained) to global threats. In particular, we have addressed the following research questions that are related to the dynamic and evolution of technological and social networks:

- Phenomenology of competing dynamical processes in networks
- Modeling human behavioral responses to the large-scale spreading of infectious diseases



Shown data HOSPITAL BEDS
Min. color Opacity 0
Max. color Opacity 0.9
DAY 122
13 MAR 2011



GLEaMviz Simulator
GLEaMviz.org

Localized Smallpox Release
- A Tutorial

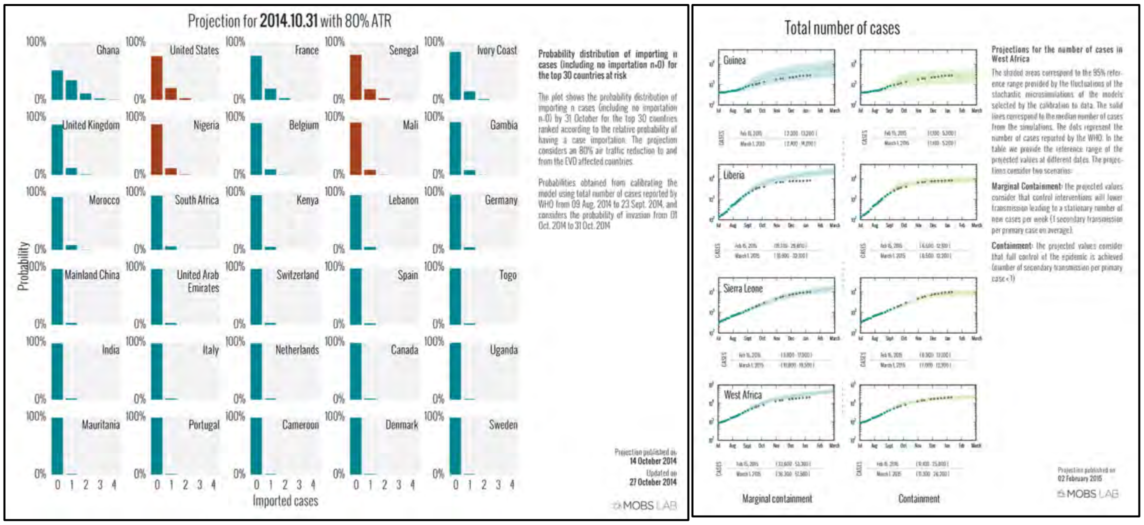
GLobal Epidemic and Mobility Modeler

www.gleamviz.org

Figure IV. : Vulnerability simulator for potential pandemic pathogens with specific plug-ins for E-WMD events.

- Tipping points in spreading and communication processes

The work in this areas have led to the developing of a coherent framework for the understanding of the global invasion threshold of pathogens in networked population, the effect of behavioral adaptation and competing spreading dynamics in contagion processes, and the development of novel algorithms for the computational modeling of global epidemic events such as pandemic and E-WMD. The results have been published in several papers, including high impact journals such as in Nature Physics [5].



Probability distribution of importing n cases (including no importation n=0) for the top 30 countries at risk

The plot shows the probability distribution of importing n cases (including no importation n=0) by 31 October for the top 30 countries ranked according to the relative probability of having a case importation. The projection considers an 80% air traffic reduction to and from the EVD affected countries.

Probabilities obtained from calibrating the model using total number of cases reported by WHO from 09 Aug. 2014 to 23 Sept. 2014, and considers the probability of invasion from 01 Oct. 2014 to 31 Oct. 2014

Projections for the number of cases in West Africa

The shaded areas correspond to the 95% reference range provided by the fluctuations of the stochastic microsimulations of the models selected by the calibration to data. The solid lines correspond to the median number of cases from the simulations. The dots represent the number of cases reported by the WHO. In the table we provide the reference range of the projected values at different dates. The projections consider two scenarios:

Marginal Containment: the projected values consider that control interventions will lower transmission leading to a stationary number of new cases per week (1 secondary transmission per primary case on average).

Containment: the projected values consider that full control of the epidemic is achieved (number of secondary transmission per primary case < 1)

Figure V.: Left- Projections for the number of possible importation of ebola cases in top risk countries. Right- projections for the evolution of the outbreak in the West-African countries.

Another main goal targeted by our project was to address the problem of the non-local effects and response induced by localized WMD attacks in complex multiscale networks. In this area we focused on the level of threat for a specific country in the case of a bioterrorist attack outside its borders, namely, a secondary E-WMD threat. We were interested in the quantitative assessment of the risk of the internationalization of the outbreak and the number of countries affected. In particular, we assumed that the bio-WMD event is localized in an area where controllability cannot be achieved and develop appropriate strategies that focus on the long-range effect of the event and halt the spreading at the global level. The results obtained show that biological targeted attacks on a single country or city can result in a global threat. In particular we were in the position to quantitatively describe the level of threat for USA in the case of a bioterrorist attack outside its borders, namely, a European country such as UK or France (secondary E-WMD threat). This work opens the way to more refined study defining novel strategies for the containment of non-local intentional/accidental release of potential pandemic pathogens (i.e. effect on the US soil of a bioterrorist attack in a foreign country). The results of our work have been published in interdisciplinary journals [Scientific Reports 3, 810 (2013); Scientific Reports 1, 62 (2011); BMC I. Diseases 11, 37 (2011)] and we have also implemented plug-ins and tutorial for the use of the GLEAMviz simulator in the study of smallpox-like E-WMD events. This computational tool is publicly available online (www.gleamviz.org) and consists of a simulation component and a suite of analysis tools.

The theoretical understanding of socio-technical networks properties and dynamics and the modeling tools developed during the project are of general interest and applications in the context of biological threats, as proven by work done during the project for two major epidemics: the H1N1-2009 pandemic, and the Ebola virus West Africa epidemic.

During both events the results achieved during the project have been put to work to provide real-time analysis of the epidemic unfolding and the risk of international spread of the pathogen. Interestingly, these applications allowed us to advance on specific points of the planned work by providing real-world scenarios and validation data. In particular, we capitalized on the work done on the non-local effects and response induced by localized WMD attacks to provide real-time analysis of the 2014 Ebola epidemic in West Africa and the associated risk of international spread. In order to assess the likelihood of the international diffusion of the epidemic we have used a specific Ebola Virus Disease (EVD) model that considers explicitly that Ebola transmissions occur in the general community, in hospital settings, and during funeral rites. This model has been used to provide estimates for the growth of the epidemic in West Africa, the probability and expected number of Ebola virus disease case importation in countries across the world. Notably, the results obtained on the EVD epidemic has been communicated to a number of governmental and non-governmental agencies and used by the World Bank in their reports on the economic impact of the 2014 Ebola epidemic for West Africa[1]. In order to improve the spatial resolution and realism of the within-country analysis we have also a spatial agent based model based on a synthetic structured population of Liberia generating a large-scale networks of the population contacts. We used the model to project the spatiotemporal spreading of the disease and to disentangle the effect of different interventions including Ebola treatment unit availability, and safe burial procedures. The results, published in PLOS Curr. Outbreaks, Euroseurveillance and Lancet Infectious Diseases, quantified the differences in the results obtained by the use of different mobility proxies and shown that the methodologies developed during the project can be used in real time situation –like the EVD outbreak – to provide quantitative indications on the local and global spreading of biological threats.

[1] World Bank. 2014. The economic impact of the 2014 Ebola epidemic: short and medium term estimates for West Africa. Washington, DC: World Bank Group. http://documents.worldbank.org/curated/en/2014/10/20270083/economic-impact-2014-ebola-epidemic-short-medium-term-estimates-west-africa

II. Dissemination, educational and outreach activities

As this was a highly interdisciplinary project we have placed special effort in disseminating the results achieved during the grant both within the group of our peers and outside the circle of technical practitioners. In terms of publications we have also been targeting broad readership journals to ensure dissemination across a wide range of scientific communities. Along with the publication outlet we have made a specific effort to be present our results in annual top-ranked conferences and other events that have gathered international experts in the area of complex networks.

Presentations given (selected):

- Modeling the Spread and Control of Ebola in West Africa: Predictions, Surveillance and Interventions, Georgia tech, Atlanta January 21-23, 2015 (A.Vespignani, Invited panelist and presenter).
- SIAM Conference on Applications of Dynamical Systems, May 17-21 (2015), Snowbird, Utah (Z. Toroczkai, invited speaker).
- “Spatial modeling of Ebola Virus Disease” London School of Tropical Medicine, London, England, Feb. 16-18 2015 (A. Vespignani invited talk)
- SIAM Conference on Computational Science, Mar 14-18 (2015), Salt Lake City, Utah (Z. Toroczkai, invited speaker).
- "Human mobility and the spreading of infectious disease", NETMOB 2015, MIT Media Lab, Boston MA, April 8th (2015) (A. Vespignani keynote speaker)
- Center for Nonlinear Studies (CNLS) Colloquium, Los Alamos National Laboratory, NM, Feb 10, 2015. (Z. Toroczkai)
- “Transportation Networks and the spread of emerging infectious diseases”, European Conference on Complex Systems, September 29, 2014, Comptrans workshop, Lucca Italy (A.Vespignani, keynote speaker).
- European Conference on Complex Systems (ECCS'14). Institue for Advanced Studies, Lucca, Italy, Sep 24, 2014. (Z. Toroczkai, invited speaker)
- International Conference on Statistical Physics, Rhodes, Greece, Jul 7-14, 2014. (Z. Toroczkai, invited speaker)
- Next Generation Surveillance for the Next Pandemic, Santa-Fe Institute, Santa Fe, 19-22 May 2014 (A. Vespignani, invited talk)
- CompleNet 2014, Bologna, Italy, March 12-14, 2014 (A. Vespignani keynote speaker)
- 5th Workshop on information in Networks, NYU, October 4-5 2013 (A. Vespignani plenary speaker)

- Digital Surveillance Meeting, CDC, Atlanta, August 6 (2013) (A. Vespignani invited talk)
- ARS'13 International Workshop “Networks in space and time: Models, Data collection and Applications”, Rome Italy, June 20-22, 2013. (A. Vespignani Keynote talk)
- IWSOS, 7th International Workshop on Self-Organizing Systems, Palma de Mallorca, Spain, May 9-10, 2013. (A. Vespignani Keynote talk)
- SIAM Conference on Applications of Dynamical Systems, May 19-23, 2013, Snowbird Utah. Invited presentation by Z. Toroczkai on modeling network traffic and damage consequences.
- APS March Meeting, Invited Session: Statistical Physics for Systemic Risk and Infrastructural Interdependencies, Baltimore, Maryland, March 18–22, 2013 (A. Vespignani invited talk)
- AAAS 2013 annual Meeting, Predictability: from physical to Data Sciences, Boston, February 16, 2013 (A. Vespignani invited talk)
- 5th Annual Global Empowerment Meeting (GEM12), Harvard University, Boston, October 24 - 25, 2012 (A. Vespignani invited talk)
- MAPCON12. Max-Planck Institut fur Physik Komplexer Systeme, International Workshop: Mathematical Physics of Complex Networks: From Graph Theory to Biological Physics. May 14 - 18, 2012, Dresden, Germany. (Z Toroczkai, invited talk)
- NICO Distinguished Speaker Series, Northwestern University, Evanston, October 10 - 11, 2012 (A. Vespignani invited talk)
- BECS Seminar, Department of Biomedical Engineering and Computational Science, Aalto University School of Science. May 28, 2012. (Z Toroczkai, invited talk)
- Workshop on Information in Networks (WIN) 2011, New York City, September 28 - 29, 2012 (A. Vespignani plenary speaker)
- International Workshop on Agent-Based Models and Complex Techno-Social SystemsETH Zurich (Switzerland), July 2-4, 2012. (A.Vespignani, Keynote talk)
- NSF workshop: Workshop on Complexity Science Applied to Coupled Infrastructure Systems (InfraPlex), Martha's Vineyard, June 3-4, 2012. (A. Vespignani, Invited Talk).
- NetSci 2011, The International School and Conference on Network Science, Budapest, 6-10 June, 2011 (A. Vespignani, Invited Talk).
- CCNR Seminar, Northeastern University, March 21, 2011, Boston, MA (Z Toroczkai)
- "Frontiers in multiscale computational modeling for zoonotic epidemics", Kansas City, October 10-12, 2011 (A.Vespignani, Invited talk)
- Workshop "Data Science and Epidemiology", Center for Infectious Disease Dynamics (CIDD) at Penn State University, October 6 and 7, 2011 (A.Vespignani, Invited talk).
- Center for Scientific Computation & Mathematical Modeling (CSCAMM) Nonlinear Dynamics of Networks, April 7, 2010, University of Maryland, College Park (A.Vespignani, Invited talk).

- SAMSI Complex Networks Opening Workshop, Aug. 29 - Sep 1, 2010, NC (Z Toroczkai, invited speaker)
- Socially Coupled Systems & Informatics-Science, Computing and Decision Making in a Complex Interdependent World 2010 Conference, Old Town Alexandria, VA, July 12-14, 2010 (A. Vespignani, Invited Speaker).
- Harvard University, Beth Israel Deaconess Med. Center, seminar, June 21, 2010, Boston (Z. Toroczkai).
- Connecting the Dots Symposium, Harvard University, Cambridge, MA, October 22, 2010 (A.Vespignani, keynote speaker).
- Information Theory and Applications Workshop, Jan 01-Feb 5, 2010, University of California San Diego. (Z. Toroczkai, invited speaker).
- University of South Carolina, Mathematics Colloquium, Apr 22, 2010, Columbia. (Z. Toroczkai)

Over the years we have also made a particular outreach effort to communities outside the academic one. This includes efforts to start a dialogue with policy makers and stakeholders in governmental and non-governmental agencies interested to the resilience of complex socio-technical networks.

- Journee scientifique Investissements d'Avenir de l'Universite de Lyon, “La complexite: quels defis pour demain?", Lyon, France, Nov 5, 2014. - Public lecture by Z. Toroczkai.
- A. Vespignani, Congress of the United States briefing. ““Harnessing Complex Networks to Solve the Nation’s Grand Challenges”, Washington, DC, Sept. 10, 2013.
- International Meeting "Complex Systems Analysis: Advancing Health System Policy Design and Planning", Rockefeller Foundation, Bellagio Center, Italy, September 24 - 28, 2012 (A.Vespignani, Panelist).
- 39th General Assembly of the Geneva Association, Washington D.C. June 6-9, 2012 (A. Vespignani, Invited presentation)
- First Global Symposium on Health Systems Research, Session on complex systems, Montreaux, Switzerland November 16-19, 2010 (A.Vespignani, Panelist).
- Influenza H1N1 modeling working group meeting, European Center for Disease Control ECDC, Stockholm, 19 October 2010 (A.Vespignani, Panelist).

We have also made efforts to disseminate the results of the project at prominent summer schools, Advanced Master programs and other educational and training programs such as

- Modeling and Forecast of Network-Driven Contagion Processes”, American Society for Microbiology Conference for Undergraduate Educators (ASMCUE), Danvers MA, May 15-18, 2014 (A. Vespignani Invited Talk).

- The 2015 Summer Institute in Statistics and Modeling in Infectious Diseases (SISMID), University of Washington in Seattle, Seattle July 12-15, 2015. (A.Vespignani, Instructor of the module "stochastic simulation methods)
- The 2014 Summer Institute in Statistics and Modeling in Infectious Diseases (SISMID), University of Washington in Seattle, Seattle July 15-18, 2014. (A.Vespignani, Instructor of the module "stochastic simulation methods)
- NetSci2013, International School and Conference on Network Science, June 3-7, 2013, Copenhagen, Denmark. Selected talk by student Y. Ren on the extended radiation model. Advisor Z. Toroczkai.
- The 2013 Summer Institute in Statistics and Modeling in Infectious Diseases (SISMID), Seattle July 16-19, 2013, University of Washington in Seattle. (A.Vespignani, Instructor of the module "stochastic simulation methods)
- 2012 Summer Institute in Statistics and Modeling in Infectious Diseases (SISMID), 9-25 July 2012, University of Washington in Seattle. (A.Vespignani, Instructor of the module "stochastic simulation methods)
- The 2011 Summer Institute in Statistics and Modeling in Infectious Diseases (SISMID), June 13-29, 2011, University of Washington in Seattle. (A.Vespignani, Instructor of the module "stochastic simulation methods)
- Master's class: From Data to Decision, British Columbia Center for Disease Control, Vancouver, May 30-June 1st, 2012 (A.Vespignani, presenting a module on epidemic modeling).
- NetSci 2011. The International School and Conference on Network Science. School lecturer/instructor and speaker (Z. Toroczkai). 6-10 Jun 2011, Budapest, Hungary

Theses, dissertations

During the life of the project we have supported a number of graduate and post doctoral students who have benefited of the training and professional development offered by working on the project. Students and post doctoral collaborators have acquired new skills in a wide range of areas, including of data mining, data visualization, computational methods, mathematics, Network Science, Statistical mechanics and High Performance computing. Domain specific knowledge have been provided to students in the areas of complex networks, contagion processes, reaction-diffusion systems. Below we provide a list of graduate students and post doctoral associates funded by the project and their current position:

- Dr. Y. Ren (Physics PhD, ND, 2015): "Betweenness Centrality and its Applications from Modeling Traffic Flows to Network Community Detection". Advisor: Z. Toroczkai, currently postdoc at Virginia Bioinformatics Institute

- Dr. Bruno Goncalves, currently Professor at Marseille University, France and Data Science fellow at NYU's Center for Data Science. Advisor: A. Vespignani
- Dr. S. Pandit (postdoctoral associate at ND), 2011. Current position: industry. Advisor: Z. Toroczkai.
- Dr. Fabio Ciulla, (Physics PhD, NEU, 2015): "Diffusion Processes in Geographical Networks", currently Data Scientist at QUID, San Francisco Ca. Advisor: A. Vespignani.
- Dr. Maria Ercsey-Ravasz (postdoctoral associate at ND), currently faculty at Babes-Bolyai University.
- Dr. Hao Hu, (Physics PhD, IU, 2010): "Reaction-Diffusion Process and the Modeling of the Spatial Spread of Infectious Diseases". Currently Sr. Research Manager at the Institute for Disease Modeling, Seattle. Advisor: A. Vespignani
- M. Varga (physics graduate student, ND) expected to graduate in 2015. Advisor: Z. Toroczkai.
- Dr. Ana Pastore Y Piontti, currently Research Associate at Northeastern University. Advisor: A. Vespignani
- Dr. Qian Zhang, currently Research Associate at Northeastern University. Advisor: A. Vespignani
- A. Strathman (Physics PhD, ND, 2013): "Applications of statistical mechanics to the modeling of social networks." Advisor: Z. Toroczkai
- H. Kim (Physics PhD, ND, 2011). Currently research at University of Arizona. Advisor: Z. Toroczkai.
- A. Asztalos (Physics PhD, ND, 2010). Currently scientific programmer at NIH. Advisor: Z. Toroczkai.

III. Publications

Below we list the publications and their abstracts in inverse chronological order that resulted from activity from this grant.

[1] S. Horvát, É. Czabarka, and Z. Toroczkai, "Reducing Degeneracy in Maximum Entropy Models of Networks," *Phys. Rev. Lett.*, vol. 114, no. 15, p. 158701, **2015**.

> Abstract: Based on Jaynes's maximum entropy principle, exponential random graphs provide a family of principled models that allow the prediction of network properties as constrained by empirical data (observables). However, their use is often hindered by the degeneracy problem characterized by spontaneous symmetry breaking, where predictions fail. Here we show that degeneracy appears when the corresponding density of states function is not log-concave, which is typically the consequence of nonlinear relationships between the constraining observables. Exploiting these nonlinear relationships here we propose a solution to the degeneracy problem for a large class of systems via transformations that render the density of states function log-concave. The effectiveness of the method is demonstrated on examples.

[2] K. E. Bassler and T. Z. Genio, Charo I Del, Erdős Péter L, Miklós István, "Exact sampling of graphs with prescribed degree correlations," *New J. Phys.*, vol. 17, no. 8, p. 083052, **2015**.

> Abstract: Many real-world networks exhibit correlations between the node degrees. For instance, in social networks nodes tend to connect to nodes of similar degree and conversely, in biological and technological networks, high-degree nodes tend to be linked with low-degree nodes. Degree correlations also affect the dynamics of processes supported by a network structure, such as the spread of opinions or epidemics. The proper modelling of these systems, i.e., without uncontrolled biases, requires the sampling of networks with a specified set of constraints. We present a solution to the sampling problem when the constraints imposed are the degree correlations. In particular, we develop an exact method to construct and sample graphs with a specified joint-degree matrix, which is a matrix providing the number of edges between all the sets of nodes of a given degree, for all degrees, thus completely specifying all pairwise degree correlations, and additionally, the degree sequence itself. Our algorithm always produces independent samples without backtracking. The complexity of the graph construction algorithm is $\mathcal{O}(NM)$ where N is the number of nodes and M is the number of edges.

[3] P. Erdös, I. Miklós, and Z. Toroczkai, "A Decomposition Based Proof for Fast Mixing of a Markov Chain over Balanced Realizations of a Joint Degree Matrix," SIAM J. Discret. Math., vol. 29, no. 1, pp. 481–499, Jan. **2015**.

Abstract: A joint degree matrix (JDM) specifies the number of connections between nodes of given degrees in a graph, for all degree pairs, and uniquely determines the degree sequence of the graph. We consider the space of all balanced realizations of an arbitrary JDM, realizations in which the links between any two fixed-degree groups of nodes are placed as uniformly as possible. We prove that a swap Markov chain Monte Carlo algorithm in the space of all balanced realizations of an arbitrary graphical JDM mixes rapidly, i.e., the relaxation time of the chain is bounded from above by a polynomial in the number of nodes n. To prove fast mixing, we first prove a general factorization theorem similar to the Martin–Randall method for disjoint decompositions (partitions). This theorem can be used to bound from below the spectral gap with the help of fast mixing subchains within every partition and a bound on an auxiliary Markov chain between the partitions. Our proof of the general factorization theorem is direct and uses conductance based methods (Cheeger inequality).

[4] C. Orsini, M. M. Dankulov, P. Colomer-de-Simón, A. Jamakovic, P. Mahadevan, A. Vahdat, K. E. Bassler, Z. Toroczkai, M. Boguñá, G. Caldarelli, S. Fortunato, and D. Krioukov, "Quantifying randomness in real networks," *Nature Communications*, vol. 6, no. May, p. 8627, **2015**.

Abstract: Represented as graphs, real networks are intricate combinations of order and disorder. Fixing some of the structural properties of network models to their values observed in real networks, many other properties appear as statistical consequences of these fixed observables, plus randomness in other respects. Here we employ the dk-series, a complete set of basic characteristics of the network structure, to study the statistical dependencies between different network properties. We consider six real networks—the Internet, US airport network, human protein interactions, techno-social web of trust, English word network, and an fMRI map of the human brain—and find that many important local and global structural properties of these networks are closely reproduced by dk-random graphs whose degree distributions, degree correlations and clustering are as in the corresponding real network. We discuss important conceptual, methodological, and practical implications of this evaluation of network randomness, and release software to generate dk-random graphs.

[5] C. Poletto, S. Meloni, A. Van Metre, V. Colizza, Y. Moreno, and A. Vespignani, "Characterising two-pathogen competition in spatially structured environments," *Sci. Rep.*, vol. 5, p. 7895, **2015**.

Abstract: Different pathogens spreading in the same host population often generate complex co-circulation dynamics because of the many possible interactions between the pathogens and the host immune system, the host life cycle, and the space structure of the population. Here we focus on the competition between two acute infections and we address the role of host mobility and cross-immunity in shaping possible dominance/co-dominance regimes. Host mobility is modelled as a network of traveling flows connecting nodes of a metapopulation, and the two-pathogen dynamics is simulated with a stochastic mechanistic approach. Results depict a complex scenario where, according to the relation among the epidemiological parameters of the two pathogens, mobility can either be non-influential for the competition dynamics or play a critical role in selecting the dominant pathogen. The characterization of the parameter space can be explained in terms of the trade-off between pathogen's spreading velocity and its ability to diffuse in a sparse environment. Variations in the cross-immunity level induce a transition between presence and absence of competition. The present study disentangles the role of the relevant biological and ecological factors in the competition dynamics, and provides relevant insights into the spatial ecology of infectious diseases.

[6] S. Merler, M. Ajelli, L. Fumanelli, M. F. C. Gomes, A. P. Y. Piontti, L. Rossi, D. L. Chao, I. M. J. Longini, M. E. Halloran, and A. Vespignani, "Spatiotemporal spread of the 2014 outbreak of Ebola virus disease in Liberia and the effectiveness of non-pharmaceutical interventions: a computational modelling analysis.," *Lancet. Infect. Dis.*, vol. 3099, no. 14, pp. 1–8, **2015**.

Abstract: BACKGROUND: The 2014 epidemic of Ebola virus disease in parts of west Africa defines an unprecedented health threat. We developed a model of Ebola virus transmission that integrates detailed geographical and demographic data from Liberia to overcome the limitations of non-spatial approaches in projecting the disease dynamics and assessing non-pharmaceutical control interventions. METHODS: We modelled the movements of individuals, including patients not infected with Ebola virus, seeking assistance in health-care facilities, the movements of individuals taking care of patients infected with Ebola virus not admitted to hospital, and the attendance of funerals. Individuals were grouped into randomly assigned households (size based on Demographic Health Survey data) that were geographically placed to match population density estimates on a grid of 3157 cells covering the country. The spatial agent-based model was calibrated with a Markov chain Monte Carlo approach. The model was used to estimate Ebola virus transmission parameters and investigate the effectiveness of interventions such as availability of Ebola treatment units, safe burials procedures, and household protection kits. FINDINGS: Up to Aug 16, 2014, we estimated that 38.3% of infections (95% CI 17.4-76.4) were acquired in hospitals, 30.7% (14.1-46.4) in households, and 8.6% (3.2-11.8) while participating in funerals. We noted that the movement and mixing, in hospitals at the early stage of the epidemic, of patients infected with Ebola virus and those not infected was a sufficient driver of the reported pattern of spatial spread. The subsequent decrease of incidence at country and county level is attributable to the increasing availability of Ebola treatment units (which in turn contributed to drastically decreased hospital transmission), safe burials, and distribution of household protection kits. INTERPRETATION: The model allows assessment of intervention options and the understanding of their role in the decrease in incidence reported since Sept 7, 2014. High-quality data (eg, to estimate household secondary attack rate, contact patterns within hospitals, and effects of ongoing interventions) are needed to reduce uncertainty in model estimates. FUNDING: US Defense Threat Reduction Agency, US National Institutes of Health.

[7] Y. Ren, M. Ercsey-Ravasz, P. Wang, M. C. González, and Z. Toroczkai, "Predicting commuter flows in spatial networks using a radiation model based on temporal ranges." *Nature Communications*, vol. 5, p. 5347, **2014**.

Abstract: Understanding network flows such as commuter traffic in large transportation networks is an ongoing challenge due to the complex nature of the transportation infrastructure and human mobility. Here we show a first-principles based method for traffic prediction using a cost-based generalization of the radiation model for human mobility, coupled with a cost-minimizing algorithm for efficient distribution of the mobility fluxes through the network. Using US census and highway traffic data, we show that traffic can efficiently and accurately be computed from a range-limited, network betweenness type calculation. The model based on travel time costs captures the log-normal distribution of the traffic and attains a high Pearson correlation coefficient (0.75) when compared with real traffic. Because of its principled nature, this method can inform many applications related to human mobility driven flows in spatial networks, ranging from transportation, through urban planning to mitigation of the effects of catastrophic events.

[8] F. Ciulla, N. Perra, A. Baronchelli, and A. Vespignani, "Damage detection via shortest-path network sampling," *Phys. Rev. E*, vol. 89, no. 5, p. 052816, **2014**.

Abstract: Large networked systems are constantly exposed to local damages and failures that can alter their functionality. The knowledge of the structure of these systems is, however, often derived through sampling strategies whose effectiveness at damage detection has not been thoroughly investigated so far. Here, we study the performance of shortest-path sampling for damage detection in large-scale networks. We define appropriate metrics to characterize the sampling process before and after the damage, providing statistical estimates for the status of nodes (damaged, not damaged). The proposed methodology is flexible and allows tuning the trade-off between the accuracy of the damage detection and the number of probes used to sample the network. We test and measure the efficiency of our approach considering both synthetic and real networks data. Remarkably, in all of the systems studied, the number of correctly identified damaged nodes exceeds the number of false positives, allowing us to uncover the damage precisely.

[9] M. F. C. Gomes, A. Pastore y Piotti, L. Rossi, D. Chao, I. Longini, and M. E. Halloran, "Assessing the International Spreading Risk Associated with the 2014 West African Ebola Outbreak," *PLOS Curr. Outbreaks*, pp. 1–22, **2014**.

Abstract: Background: The 2014 West African Ebola Outbreak is so far the largest and deadliest recorded in history. The affected countries, Sierra Leone, Guinea, Liberia, and

Nigeria, have been struggling to contain and to mitigate the outbreak. The ongoing rise in confirmed and suspected cases, 2615 as of 20 August 2014, is considered to increase the risk of international dissemination, especially because the epidemic is now affecting cities with major commercial airports. Method: We use the Global Epidemic and Mobility Model to generate stochastic, individual based simulations of epidemic spread worldwide, yielding, among other measures, the incidence and seeding events at a daily resolution for 3,362 subpopulations in 220 countries. The mobility model integrates daily airline passenger traffic worldwide and the disease model includes the community, hospital, and burial transmission dynamic. We use a multimodel inference approach calibrated on data from 6 July to the date of 9 August 2014. The estimates obtained were used to generate a 3-month ensemble forecast that provides quantitative estimates of the local transmission of Ebola virus disease in West Africa and the probability of international spread if the containment measures are not successful at curtailing the outbreak. Results: We model the short-term growth rate of the disease in the affected West African countries and estimate the basic reproductive number to be in the range $1.5 - 2.0$ (interval at the 1/10 relative likelihood). We simulated the international spreading of the outbreak and provide the estimate for the probability of Ebola virus disease case importation in countries across the world. Results indicate that the short-term (3 and 6 weeks) probability of international spread outside the African region is small, but not negligible. The extension of the outbreak is more likely occurring in African countries, increasing the risk of international dissemination on a longer time scale.

[10] N. T. Markov, M. M. Ercsey-Ravasz, A. R. Ribeiro Gomes, C. Lamy, L. Magrou, J. Vezoli, P. Misery, A. Falchier, R. Quilodran, M. A. Gariel, J. Sallet, R. Gamanut, C. Huissoud, S. Clavagnier, P. Giroud, D. Sappey-Marinier, P. Barone, C. Dehay, Z. Toroczkai, K. Knoblauch, D. C. Van Essen, and H. Kennedy, "A Weighted and Directed Interareal Connectivity Matrix for Macaque Cerebral Cortex," *Cereb. Cortex*, vol. 24, no. 1, pp. 17–36, **2014**.

Abstract: Retrograde tracer injections in 29 of the 91 areas of the macaque cerebral cortex revealed 1,615 interareal pathways, a third of which have not previously been reported. A weight index (extrinsic fraction of labeled neurons [FLNe]) was determined for each area-to-area pathway. Newly found projections were weaker on average compared with the known projections; nevertheless, the 2 sets of pathways had extensively overlapping weight distributions. Repeat injections across individuals revealed modest FLNe variability given the range of FLNe values (standard deviation <1 log unit, range 5 log units). The connectivity profile for each area conformed to a lognormal distribution, where a majority of projections are moderate or weak in strength. In the $G_{29\times29}$ interareal subgraph, two-thirds of the connections that can exist do exist. Analysis of the smallest set of areas that collects links from all 91 nodes of the $G_{29\times91}$ subgraph (dominating set analysis) confirms the dense (66%) structure of the cortical matrix. The $G_{29\times29}$ subgraph suggests an unexpectedly high incidence of unidirectional links. The directed and weighted $G_{29\times91}$

connectivity matrix for the macaque will be valuable for comparison with connectivity analyses in other species, including humans. It will also inform future modeling studies that explore the regularities of cortical networks.

[11] A. Pastore Y Piontti, M. F. D. C. Gomes, N. Samay, N. Perra, and A. Vespignani, "The infection tree of global epidemics," *Netw. Sci.*, vol. 2, no. 01, pp. 132–137, **2014**.

[12] C. Poletto, M. Gomes, A. Pastore y Piontti, L. Rossi, L. Bioglio, D. Chao, I. Longini, M. Halloran, V. Colizza, and A. Vespignani, "Assessing the impact of travel restrictions on international spread of the 2014 West African Ebola epidemic," *Eurosurveillance*, vol. 19, no. 42, p. 20936, **2014**.

Abstract: The quick spread of an Ebola outbreak in West Africa has led a number of countries and airline companies to issue travel bans to the affected areas. Considering data up to 31 Aug 2014, we assess the impact of the resulting traffic reductions with detailed numerical simulations of the international spread of the epidemic. Traffic reductions are shown to delay by only a few weeks the risk that the outbreak extends to new countries.

[13] C. Poletto, S. Meloni, V. Colizza, Y. Moreno, and A. Vespignani, "Host Mobility Drives Pathogen Competition in Spatially Structured Populations," *PLoS Comput Biol*, vol. 9, no. 8, p. e1003169, **2013**.

Abstract: Author SummaryWhen multiple infectious agents circulate in a given population of hosts, they interact for the exploitation of susceptible hosts aimed at pathogen survival and maintenance. Such interaction is ruled by the combination of different mechanisms related to the biology of host-pathogen interaction, environmental conditions and host demography and behavior. We focus on pathogen competition and we investigate whether the mobility of hosts in a spatially structured environment can act as a selective driver for pathogen circulation. We use mathematical and computational models for disease transmission between hosts and for the mobility of hosts to study the competition between two pathogens providing each other full cross-immunity after infection. Depending on the rate of migration of hosts, competition results in the dominance of either one of the pathogens at the spatial level – though the two infectious agents are characterized by the same invasion potential at the single population scale – or cocirculation of both. These results highlight the importance of explicitly accounting for the spatial scale and for the different time scales involved (i.e. host mobility and spreading dynamics of the two pathogens) in the study of host-multipathogen systems.

[14] M. Ercsey-Ravasz, N. T. Markov, C. Lamy, D. C. Van Essen, K. Knoblauch, Z. Toroczkai, and H. Kennedy, "A Predictive Network Model of Cerebral Cortical Connectivity Based on a Distance Rule," *Neuron*, vol. 80, no. 1, pp. 184–197, **2013**.

Abstract: Recent advances in neuroscience have engendered interest in large-scale brain networks. Using a consistent database of cortico-cortical connectivity, generated from hemisphere-wide, retrograde tracing experiments in the macaque, we analyzed interareal weights and distances to reveal an important organizational principle of brain connectivity. Using appropriate graph theoretical measures, we show that although very dense (66%), the interareal network has strong structural specificity. Connection weights exhibit a heavy-tailed lognormal distribution spanning five orders of magnitude and conform to a distance rule reflecting exponential decay with interareal separation. A single-parameter random graph model based on this rule predicts numerous features of the cortical network: (1) the existence of a network core and the distribution of cliques, (2) global and local binary properties, (3) global and local weight-based communication efficiencies modeled as network conductance, and (4) overall wire-length minimization. These findings underscore the importanceof distance and weight-based heterogeneity in cortical architecture and processing.

[15] B. Gonçalves, D. Balcan, and A. Vespignani, "Human mobility and the worldwide impact of intentional localized highly pathogenic virus release.," *Sci. Rep.*, vol. 3, p. 810, **2013**.

Abstract: The threat of bioterrorism and the possibility of accidental release have spawned a growth of interest in modeling the course of the release of a highly pathogenic agent. Studies focused on strategies to contain local outbreaks after their detection show that timely interventions with vaccination and contact tracing are able to halt transmission. However, such studies do not consider the effects of human mobility patterns. Using a large-scale structured metapopulation model to simulate the global spread of smallpox after an intentional release event, we show that index cases and potential outbreaks can occur in different continents even before the detection of the pathogen release. These results have two major implications: i) intentional release of a highly pathogenic agent within a country will have global effects; ii) the release event may trigger outbreaks in countries lacking the health infrastructure necessary for effective containment. The presented study provides data with potential uses in defining contingency plans at the National and International level.

[16] N. T. Markov, M. Ercsey-Ravasz, C. Lamy, A. R. Ribeiro Gomes, L. Magrou, P. Misery, P. Giroud, P. Barone, C. Dehay, Z. Toroczkai, K. Knoblauch, D. C. Van Essen, and H. Kennedy, "The role of long-range connections on the specificity of the macaque interareal cortical network.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, no. 13, pp. 5187–92, **2013**.

Abstract: We investigated the influence of interareal distance on connectivity patterns in a database obtained from the injection of retrograde tracers in 29 areas distributed over six regions (occipital, temporal, parietal, frontal, prefrontal, and limbic). One-third of the 1,615 pathways projecting to the 29 target areas were reported only recently and deemed new-found projections (NFPs). NFPs are predominantly long-range, low-weight connections. A minimum dominating set analysis (a graph theoretic measure) shows that NFPs play a major role in globalizing input to small groups of areas. Randomization tests show that (i) NFPs make important contributions to the specificity of the connectivity profile of individual cortical areas, and (ii) NFPs share key properties with known connections at the same distance. We developed a similarity index, which shows that intraregion similarity is high, whereas the interregion similarity declines with distance. For area pairs, there is a steep decline with distance in the similarity and probability of being connected. Nevertheless, the present findings reveal an unexpected binary specificity despite the high density (66%) of the cortical graph. This specificity is made possible because connections are largely concentrated over short distances. These findings emphasize the importance of long-distance connections in the connectivity profile of an area. We demonstrate that long-distance connections are particularly prevalent for prefrontal areas, where they may play a prominent role in large scale communication and information integration.

[17] S. Merler, M. Ajelli, L. Fumanelli, and A. Vespignani, "Containing the accidental laboratory escape of potential pandemic influenza viruses.," *BMC Med.*, vol. 11, p. 252, **2013**.

Abstract: BACKGROUND: The recent work on the modified H5N1 has stirred an intense debate on the risk associated with the accidental release from biosafety laboratory of potential pandemic pathogens. Here, we assess the risk that the accidental escape of a novel transmissible influenza strain would not be contained in the local community. METHODS: We develop here a detailed agent-based model that specifically considers laboratory workers and their contacts in microsimulations of the epidemic onset. We consider the following non-pharmaceutical interventions: isolation of the laboratory, laboratory workers' household quarantine, contact tracing of cases and subsequent household quarantine of identified secondary cases, and school and workplace closure both preventive and reactive. RESULTS: Model simulations suggest that there is a non-negligible probability (5% to 15%), strongly dependent on reproduction number and probability of developing clinical symptoms, that the escape event is not detected at all. We find that the containment depends on the timely implementation of non-pharmaceutical interventions and contact tracing and it may be effective (>90% probability per event) only for pathogens with moderate transmissibility (reproductive number no larger than $R_0 = 1.5$). Containment depends on population density and structure as well, with a probability of giving rise to a global event that is three to five times lower in rural areas. CONCLUSIONS: Results suggest that controllability of escape events is not guaranteed and, given the rapid increase of biosafety laboratories worldwide, this poses a serious threat to human health. Our findings may be relevant to policy makers when designing adequate preparedness plans and may have important implications for determining the location of new biosafety laboratories worldwide.

[18] H. Kennedy, K. Knoblauch, and Z. Toroczkai, "Why data coherence and quality is critical for understanding interareal cortical networks," *Neuroimage*, vol. 80, pp. 37–45, **2013**.

Abstract: Numerous studies have investigated inter-areal cortical networks using either diffusion MRI or axonal tract-tracing. While both techniques have been used in non-human primates only diffusion MRI can be used in human. The advantage of axonal tract-tracing is that unlike diffusion MRI it has a high single-cell resolution, and most importantly gives the laminar origins and terminations of inter-areal pathways. It, therefore, can be used to obtain the weighted and directed cortical graph. Axonal tract tracing has traditionally been collated from multiple experiments in order to determine the large-scale inter-areal network. Collated data of this kind present numerous problems due to lack of coherence across studies and incomplete exploitation. We have therefore developed a consistent data base which uses standardized experimental and parcellation procedures across brains. Here we review our recent publications analyzing the consistent database obtained from retrograde tracer injections in 29 cortical areas in a parcellation of 91 areas of the macaque cortex. Compared to collated data, our results show that the cortical graph is dense. Density is a graph theoretic measure, and refers to the number of observed connections in a square matrix expressed as a percentage of the possible connections. In our database 66% of the connections that can exist do exist which is considerably higher than the graph densities reported in studies using collated data (7–32%). The consistent data base reports 37% more pathways than previously reported, many of which are unidirectional. This latter and unexpected property has not been reported in earlier studies. Given the high density, the resulting cortical graph shows other unexpected properties. Firstly, the binary specificity is considerably higher than expected. As we show, this property is a consequence of the inter-areal connection probability declining with distance. Secondly, small groups of areas are found to receive high numbers of inputs. This is termed a high domination and is analyzed by a graph theoretic procedure known as a minimum dominating set analysis. We discuss these findings with respect to the long-distance connections, over half of which were previously not reported. These so called new found projections display high specificities and play an important integration role across large regions. It is to be expected that the future examination of the 62 remaining areas will disclose further levels of complexity and enable construction of a weighted directed graph revealing the hierarchical complexity of the cortex.

[19] C. Wang, O. Lizardo, D. Hachen, A. Strathman, Z. Toroczkai, and N. V. Chawla, "A dyadic reciprocity index for repeated interaction networks," *Netw. Sci.*, vol. 1, no. 01, pp. 31–48, **2013**.

Abstract: A wide variety of networked systems in human societies are composed of repeated communications between actors. A dyadic relationship made up of repeated interactions may be reciprocal (both actors have the same probability of directing a communication attempt to the other) or non-reciprocal (one actor has a higher probability of initiating a communication attempt than the other). In this paper we propose a theoretically motivated index of reciprocity appropriate for networks formed from repeated interactions based on these probabilities. We go on to examine the distribution of reciprocity in a large-scale social network built from trace-logs of over a billion cell-phone communication events across millions of actors in a large industrialized country. We find that while most relationships tend toward reciprocity, a substantial minority of relationships exhibit large levels of non-reciprocity. This is puzzling because behavioral theories in social science predict that persons will selectively terminate non-reciprocal relationships, keeping only those that approach reciprocity. We point to two structural features of human communication behavior and relationship formation—the division of contacts into strong and weak ties and degree-based assortativity—that either help or hinder the ability of persons to obtain communicative balance in their relationships. We examine the extent to which deviations from reciprocity in the observed network are partially traceable to the operation of these countervailing tendencies.

[20] M. Ercsey-Ravasz, R. N. Lichtenwalter, N. V. Chawla, and Z. Toroczkai, "Range-limited centrality measures in complex networks," *Phys. Rev. E*, vol. 85, no. 6, p. 066103, **2012**.

Abstract: Here we present a range-limited approach to centrality measures in both non-weighted and weighted directed complex networks. We introduce an efficient method that generates for every node and every edge its betweenness centrality based on shortest paths of lengths not longer than $l = 1, \cdots, L$ in the case of non-weighted networks, and for weighted networks the corresponding quantities based on minimum weight paths with path weights not larger than $w_l = l\Delta$, $l = 1, \cdots, L = R/\Delta$. These measures provide a systematic description on the positioning importance of a node (edge) with respect to its network neighborhoods one step out, two steps out, etc., up to and including the whole network. They are more informative than traditional centrality measures, as network transport typically happens on all length scales, from transport to nearest neighbors to the farthest reaches of the network. We show that range-limited centralities obey universal scaling laws for large non-weighted networks. As the computation of traditional centrality measures is costly, this scaling behavior can be exploited to efficiently estimate centralities of nodes and edges for all ranges, including the traditional ones. The scaling behavior can also be exploited to show that the ranking top list of nodes (edges) based on their range-limited centralities quickly freezes as a function of the range, and hence the diameter-range top list can be efficiently predicted. We also show how to estimate the typical largest node-to-node distance for a network of N nodes, exploiting the afore-mentioned scaling behavior. These observations were made on model networks and on a large social network inferred from cell-phone trace logs ($\sim 5.5 \times 10^6$ nodes and $\sim 2.7 \times 10^7$ edges). Finally, we apply these concepts to efficiently detect the vulnerability backbone of a network (defined as the smallest percolating cluster of the highest betweenness nodes and edges) and illustrate the importance of weight-based centrality measures in weighted networks in detecting such backbones.

[21] D. Balcan and A. Vespignani, "Invasion threshold in structured populations with recurrent

mobility patterns," *J. Theor. Biol.*, vol. 293, pp. 87–100, **2012**.

> Abstract: In this paper we develop a framework to analyze the behavior of contagion and spreading processes in complex subpopulation networks where individuals have memory of their subpopulation of origin. We introduce a metapopulation model in which subpopulations are connected through heterogeneous fluxes of individuals. The mobility process among communities takes into account the memory of residence of individuals and is incorporated with the classical susceptible-infectious-recovered epidemic model within each subpopulation. In order to gain analytical insight into the behavior of the system we use degree-block variables describing the heterogeneity of the subpopulation network and a time-scale separation technique for the dynamics of individuals. By considering the stochastic nature of the epidemic process we obtain the explicit expression of the global epidemic invasion threshold, below which the disease dies out before reaching a macroscopic fraction of the subpopulations. This threshold is not present in continuous deterministic diffusion models and explicitly depends on the disease parameters, the mobility rates, and the properties of the coupling matrices describing the mobility across subpopulations. The results presented here take a step further in offering insight into the fundamental mechanisms controlling the spreading of infectious diseases and other contagion processes across spatially structured communities.

[22] M. Tizzoni, P. Bajardi, C. Poletto, J. J. Ramasco, D. Balcan, B. Gonçalves, N. Perra, V. Colizza, and A. Vespignani, "Real-time numerical forecast of global epidemic spreading: case study of 2009 A/H1N1pdm," *BMC Med.*, vol. 10, no. 1, p. 165, **2012**.

> Abstract: Background: Mathematical and computational models for infectious diseases are increasingly used to support public-health decisions; however, their reliability is currently under debate. Real-time forecasts of epidemic spread using data-driven models have been hindered by the technical challenges posed by parameter estimation and validation. Data gathered for the 2009 H1N1 influenza crisis represent an unprecedented opportunity to validate real-time model predictions and define the main success criteria for different approaches. Methods: We used the Global Epidemic and Mobility Model to generate stochastic simulations of epidemic spread worldwide, yielding (among other measures) the incidence and seeding events at a daily resolution for 3,362 subpopulations in 220 countries. Using a Monte Carlo Maximum Likelihood analysis, the model provided an estimate of the seasonal transmission potential during the early phase of the H1N1 pandemic and generated ensemble forecasts for the activity peaks in the northern hemisphere in the fall/winter wave. These results were validated against the real-life surveillance data collected in 48 countries, and their robustness assessed by focusing on 1) the peak timing of the pandemic; 2) the level of spatial resolution allowed by the model; and 3) the clinical attack rate and the effectiveness of the vaccine. In addition, we studied the effect of data incompleteness on the prediction reliability. Results: Real-time

predictions of the peak timing are found to be in good agreement with the empirical data, showing strong robustness to data that may not be accessible in real time (such as pre-exposure immunity and adherence to vaccination campaigns), but that affect the predictions for the attack rates. The timing and spatial unfolding of the pandemic are critically sensitive to the level of mobility data integrated into the model. Conclusions: Our results show that large-scale models can be used to provide valuable real-time forecasts of influenza spreading, but they require high-performance computing. The quality of the forecast depends on the level of data integration, thus stressing the need for high-quality data in population-based models, and of progressive updates of validated available empirical knowledge to inform these models.

[23] H. Kim, C. I. Del Genio, K. E. Bassler, and Z. Toroczkai, "Constructing and sampling directed graphs with given degree sequences," *New J. Phys.*, vol. 14, no. 2, p. 023012, **2012**.

Abstract: The interactions between the components of complex networks are often directed. Proper modeling of such systems frequently requires the construction of ensembles of digraphs with a given sequence of in- and out-degrees. As the number of simple labeled graphs with a given degree sequence is typically very large even for short sequences, sampling methods are needed for statistical studies. Currently, there are two main classes of methods that generate samples. One of the existing methods first generates a restricted class of graphs and then uses a Markov chain Monte-Carlo algorithm based on edge swaps to generate other realizations. As the mixing time of this process is still unknown, the independence of the samples is not well controlled. The other class of methods is based on the configuration model that may lead to unacceptably many sample rejections due to self-loops and multiple edges. Here we present an algorithm that can directly construct all possible realizations of a given bi-degree sequence by simple digraphs. Our method is rejection-free, guarantees the independence of the constructed samples and provides their weight. The weights can then be used to compute statistical averages of network observables as if they were obtained from uniformly distributed sampling or from any other chosen distribution.

[24] M. Ercsey-Ravasz, Z. Toroczkai, Z. Lakner, and J. Baranyi, "Complexity of the international agro-food trade network and its impact on food safety.," *PLoS One*, vol. 7, no. 5, p. e37810, **2012**.

Abstract: With the world's population now in excess of 7 billion, it is vital to ensure the chemical and microbiological safety of our food, while maintaining the sustainability of its production, distribution and trade. Using UN databases, here we show that the international agro-food trade network (IFTN), with nodes and edges representing countries and import export fluxes, respectively, has evolved into a highly heterogeneous, complex supply-chain network. Seven countries form the core of the IFTN, with high values of betweenness centrality and each trading with over 77% of all the countries in the world. Graph theoretical analysis and a dynamic food flux model show that the IFTN provides a vehicle suitable for the fast distribution of potential contaminants but unsuitable for tracing their origin. In particular, we show that high values of node betweenness and vulnerability correlate well with recorded large food poisoning outbreaks.

[25] N. Perra, D. Balcan, B. Gonçalves, and A. Vespignani, "Towards a characterization of behavior-disease models.," *PLoS One*, vol. 6, no. 8, p. e23084, **2011**.

Abstract: The last decade saw the advent of increasingly realistic epidemic models that leverage on the availability of highly detailed census and human mobility data. Data-driven models aim at a granularity down to the level of households or single individuals. However, relatively little systematic work has been done to provide coupled behavior-disease models able to close the feedback loop between behavioral changes triggered in the population by an individual's perception of the disease spread and the actual disease spread itself. While models lacking this coupling can be extremely successful in mild epidemics, they obviously will be of limited use in situations where social disruption or behavioral alterations are induced in the population by knowledge of the disease. Here we propose a characterization of a set of prototypical mechanisms for self-initiated social distancing induced by local and non-local prevalence-based information available to individuals in the population. We characterize the effects of these mechanisms in the framework of a compartmental scheme that enlarges the basic SIR model by considering separate behavioral classes within the population. The transition of individuals in/out of behavioral classes is coupled with the spreading of the disease and provides a rich phase space with multiple epidemic peaks and tipping points. The class of models presented here can be used in the case of data-driven computational approaches to analyze scenarios of social adaptation and behavioral change.

[26] N. T. Markov, P. Misery, a. Falchier, C. Lamy, J. Vezoli, R. Quilodran, M. a. Gariel, P. Giroud, M. Ercsey-Ravasz, L. J. Pilaz, C. Huissoud, P. Barone, C. Dehay, Z. Toroczkai, D. C. Van Essen, H. Kennedy, and K. Knoblauch, "Weight consistency specifies regularities of macaque cortical networks," *Cereb. Cortex*, vol. 21, no. 6, pp. 1254–1272, **2011**.

Abstract: To what extent cortical pathways show significant weight differences and whether these differences are consistent across animals (thereby comprising robust connectivity profiles) is an important and unresolved neuroanatomical issue. Here we report a quantitative retrograde tracer analysis in the cynomolgus macaque monkey of the weight consistency of the afferents of cortical areas across brains via calculation of a weight index (fraction of labeled neurons, FLN). Injection in 8 cortical areas (3 occipital plus 5 in the other lobes) revealed a consistent pattern: small subcortical input (1.3% cumulative FLN), high local intrinsic connectivity (80% FLN), high-input form neighboring areas (15% cumulative FLN), and weak long-range corticocortical connectivity (3% cumulative FLN). Corticocortical FLN values of projections to areas V1, V2, and V4 showed heavy-tailed, lognormal distributions spanning 5 orders of magnitude that were consistent, demonstrating significant connectivity profiles. These results indicate that 1) connection weight heterogeneity plays an important role in determining cortical network specificity, 2) high investment in local projections highlights the importance of local processing, and 3) transmission of information across multiple hierarchy levels mainly involves pathways having low FLN values.

[27] W. Van den Broeck, C. Gioannini, B. Gonçalves, M. Quaggiotto, V. Colizza, and A. Vespignani, "The GLEaMviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale.," *BMC Infect. Dis.*, vol. 11, no. 1, p. 37, **2011**.

Abstract: Computational models play an increasingly important role in the assessment and control of public health crises, as demonstrated during the 2009 H1N1 influenza pandemic. Much research has been done in recent years in the development of sophisticated data-driven models for realistic computer-based simulations of infectious disease spreading. However, only a few computational tools are presently available for assessing scenarios, predicting epidemic evolutions, and managing health emergencies that can benefit a broad audience of users including policy makers and health institutions.

[28] D. Balcan and A. Vespignani, "Phase transitions in contagion processes mediated by recurrent mobility patterns," *Nat. Phys.*, vol. 7, no. 7, pp. 581–586, **2011**.

Abstract: Human mobility and activity patterns mediate contagion on many levels, including the spatial spread of infectious diseases, diffusion of rumors, and emergence of consensus. These patterns however are often dominated by specific locations and recurrent flows and poorly modelled by the random diffusive dynamics generally used to study them. Here we develop a theoretical framework to analyse contagion within a network of locations where individuals recall their geographic origins. We find a phase transition between a regime in which the contagion affects a large fraction of the system and one in which only a small fraction is affected. This transition cannot be uncovered by continuous deterministic models because of the stochastic features of the contagion process and defines an invasion threshold that depends on mobility parameters, providing guidance for controlling contagion spread by constraining mobility processes. We recover the threshold behaviour by analysing diffusion processes mediated by real human commuting data.

[29] P. Bajardi, C. Poletto, J. J. Ramasco, M. Tizzoni, V. Colizza, and A. Vespignani, "Human mobility networks, travel restrictions, and the global spread of 2009 H1N1 pandemic.," *PLoS One*, vol. 6, no. 1, p. e16591, **2011**.

Abstract: After the emergence of the H1N1 influenza in 2009, some countries responded with travel-related controls during the early stage of the outbreak in an attempt to contain or slow down its international spread. These controls along with self-imposed travel limitations contributed to a decline of about 40% in international air traffic to/from

Mexico following the international alert. However, no containment was achieved by such restrictions and the virus was able to reach pandemic proportions in a short time. When gauging the value and efficacy of mobility and travel restrictions it is crucial to rely on epidemic models that integrate the wide range of features characterizing human mobility and the many options available to public health organizations for responding to a pandemic. Here we present a comprehensive computational and theoretical study of the role of travel restrictions in halting and delaying pandemics by using a model that explicitly integrates air travel and short-range mobility data with high-resolution demographic data across the world and that is validated by the accumulation of data from the 2009 H1N1 pandemic. We explore alternative scenarios for the 2009 H1N1 pandemic by assessing the potential impact of mobility restrictions that vary with respect to their magnitude and their position in the pandemic timeline. We provide a quantitative discussion of the delay obtained by different mobility restrictions and the likelihood of containing outbreaks of infectious diseases at their source, confirming the limited value and feasibility of international travel restrictions. These results are rationalized in the theoretical framework characterizing the invasion dynamics of the epidemics at the metapopulation level.

[30] A. Vespignani, "Modelling dynamical processes in complex socio-technical systems," *Nat. Phys.*, vol. 8, no. 1, pp. 32–39, **2011**.

Abstract: In recent years the increasing availability of computer power and informatics tools has enabled the gathering of reliable data quantifying the complexity of socio-technical systems. Data-driven computational models have emerged as appropriate tools to tackle the study of dynamical phenomena as diverse as epidemic outbreaks, information spreading and Internet packet routing. These models aim at providing a rationale for understanding the emerging tipping points and nonlinear properties that often underpin the most interesting characteristics of socio-technical systems. Here, using diffusion and contagion phenomena as prototypical examples, we review some of the recent progress in modelling dynamical processes that integrates the complex features and heterogeneities of real-world systems.

[31] S. Meloni, N. Perra, A. Arenas, S. Gómez, Y. Moreno, and A. Vespignani, "Modeling human mobility responses to the large-scale spreading of infectious diseases," Sci. Rep., vol. 1, pp. 1–7, **2011**.

Abstract: Current modeling of infectious diseases allows for the study of realistic scenarios that include population heterogeneity, social structures, and mobility processes down to the individual level. The advances in the realism of epidemic description call for the explicit modeling of individual behavioral responses to the presence of disease within modeling frameworks. Here we formulate and analyze a metapopulation model that incorporates several scenarios of self-initiated behavioral changes into the mobility patterns of individuals. We find that prevalence-based travel limitations do not alter the epidemic invasion threshold. Strikingly, we observe in both synthetic and data-driven numerical simulations that when travelers decide to avoid locations with high levels of prevalence, this self-initiated behavioral change may enhance disease spreading. Our results point out that the real-time availability of information on the disease and the ensuing behavioral changes in the population may produce a negative impact on disease containment and mitigation.

[32] M. Ercsey-Ravasz and Z. Toroczkai, "Centrality scaling in large networks," *Phys. Rev. Lett.*, vol. 105, no. 3, pp. 2–5, **2010**.

Abstract: Betweenness centrality lies at the core of both transport and structural vulnerability properties of complex networks; however, it is computationally costly, and its measurement for networks with millions of nodes is nearly impossible. By introducing a multiscale decomposition of shortest paths, we show that the contributions to betweenness coming from geodesics not longer than L obey a characteristic scaling versus L, which can be used to predict the distribution of the full centralities. The method is also illustrated on a real-world social network of 5.5×10^6 nodes and 2.7×10^7 links.

[33] M. Ajelli, B. Gonçalves, D. Balcan, V. Colizza, H. Hu, J. J. Ramasco, S. Merler, and A. Vespignani, "Comparing large-scale computational approaches to epidemic modeling: agent-based versus structured metapopulation models.," *BMC Infect. Dis.*, vol. 10, p. 190, **2010**.

Abstract: Background: In recent years large-scale computational models for the realistic simulation of epidemic outbreaks have been used with increased frequency. Methodologies adapt to the scale of interest and range from very detailed agent-based models to spatially-structured metapopulation models. One major issue thus concerns to what extent the geo-temporal spreading pattern found by different modeling approaches may differ and depend on the different approximations and assumptions used. Methods: We provide for the first time a side-by-side comparison of the results obtained with a stochastic agent-based model and a structured metapopulation stochastic model for the progression of a baseline pandemic event in Italy, a large and geographically heterogeneous European country. The agent-based model is based on the explicit representation of the Italian population through highly detailed data on the socio-demographic structure. The metapopulation simulations use the GLobal Epidemic and Mobility (GLEaM) model, based on high-resolution census data worldwide, and integrating airline travel flow data with short-range human mobility patterns at the global scale. The model also considers age structure data for Italy. GLEaM and the agent-based models are synchronized in their initial conditions by using the same disease parameterization, and by defining the same importation of infected cases from international travels. Results: The results obtained show that both models provide epidemic patterns that are in very good agreement at the granularity levels accessible by both approaches, with differences in peak timing on the order of a few days. The relative difference of the epidemic size depends on the basic reproductive ratio, R0, and on the fact that the metapopulation model consistently yields a larger incidence than the agent-based model, as expected due to the differences in the structure in the intra-population contact pattern of the approaches. The age breakdown analysis shows that similar attack rates are obtained for the younger age classes. Conclusions: The good agreement between the two modeling approaches is very important for defining the tradeoff between data availability and the information provided by the models. The results we present define the possibility of hybrid models combining the agent-based and the metapopulation approaches according to the available data and computational resources.

[34] C. I. Del Genio, H. Kim, Z. Toroczkai, and K. E. Bassler, "Efficient and exact sampling of simple graphs with given arbitrary degree sequence," *PLoS One*, vol. 5, no. 4, pp. 1–8, **2010**.

Abstract: Uniform sampling from graphical realizations of a given degree sequence is a fundamental component in simulation-based measurements of network observables, with applications ranging from epidemics, through social networks to Internet modeling. Existing graph sampling methods are either link-swap based (Markov-Chain Monte Carlo algorithms) or stub-matching based (the Configuration Model). Both types are ill-controlled, with typically unknown mixing times for link-swap methods and uncontrolled rejections for the Configuration Model. Here we propose an efficient, polynomial time algorithm that generates statistically independent graph samples with a given, arbitrary, degree sequence. The algorithm provides a weight associated with each sample, allowing the observable to be measured either uniformly over the graph ensemble, or, alternatively, with a desired distribution. Unlike other algorithms, this method always produces a sample, without backtracking or rejections. Using a central limit theorem-based reasoning, we argue, that for large N, and for degree sequences admitting many realizations, the sample weights are expected to have a lognormal distribution. As examples, we apply our algorithm to generate networks with degree sequences drawn from power-law distributions and from binomial distributions.

[35] D. Balcan, B. Gonçalves, H. Hu, J. J. Ramasco, V. Colizza, and A. Vespignani, "Modeling the spatial spread of infectious diseases: The GLobal Epidemic and Mobility computational model," *J. Comput. Sci.*, vol. 1, no. 3, pp. 132–145, **2010**.

Abstract: Here we present the Global Epidemic and Mobility (GLEaM) model that integrates sociodemographic and population mobility data in a spatially structured stochastic disease approach to simulate the spread of epidemics at the worldwide scale. We discuss the flexible structure of the model that is open to the inclusion of different disease structures and local intervention policies. This makes GLEaM suitable for the computational modeling and anticipation of the spatio-temporal patterns of global epidemic spreading, the understanding of historical epidemics, the assessment of the role of human mobility in shaping global epidemics, and the analysis of mitigation and containment scenarios.

[36] R. J. Colizza V, Vespignani A, Perra N, Poletto C, Gonçalves B, Hu H, Balcan D, Paolotti D, Van den Broeck W, Tizzoni M, Bajardi P, "Estimate of ovel Influenza A / H1 1 cases in Mexico at the early stage of the pandemic with a spatially structured epidemic model," *PLOS Currents Influenza* pp. 1–9, **2010**.

<u>Abstract</u>: Determining the number of cases in an epidemic is fundamental to properly evaluate several disease features of high relevance for public health policies such as mortality, morbidity or hospitalization rates. Surveillance efforts are however incomplete especially at the early stage of an outbreak due to the ongoing learning process about the disease characteristics. An example of this is represented by the number of H1N1 influenza cases in Mexico during the first months of the current pandemic. Several estimates using backtrack calculation based on imported cases from Mexico in other countries point out that the actual number of cases was likely orders of magnitude larger than the number of confirmed cases. Realistic computational models fed with the best available estimates of the basic disease parameters can provide an ab-initio calculation of the number of cases in Mexico as other countries. Here we use the Global Epidemic and Mobility (GLEaM) model to obtain estimates of the size of the epidemic in Mexico as well as of imported cases at the end of April and beginning of May. We find that the reference range for the number of cases in Mexico on April 30th is 121,000 to 1,394,000 in good agreement with the recent estimates by Lipsitch et al. [M. Lipsitch, PloS One 4:e6895 (2009)]. The number of imported cases from Mexico in several countries is found to be in good agreement with the surveillance data.

[37] A. Asztalos and Z. Toroczkai, "Network Discovery by Generalized Random Walks," *Europhys. Lett.*, vol. 92, no. December, p. 50008, **2010**.

<u>Abstract</u>: We investigate network exploration by random walks defined via stationary and adaptive transition probabilities on large graphs. We derive an exact formula valid for arbitrary graphs and arbitrary walks with stationary transition probabilities (STP), for the average number of discovered edges as a function of time. We show that for STP walks site and edge exploration obey the same scaling $\sim n^{\lambda}$ as a function of time n. Therefore, edge exploration on graphs with many loops is always lagging compared to site exploration, the revealed graph being sparse until almost all nodes have been discovered. We then introduce the edge explorer model (EEM), which presents a novel class of adaptive walks, that perform faithful network discovery even on dense networks.

[38] A. Vespignani, "Complex networks: The fragility of interdependency.," *Nature*, vol. 464, no. 7291, pp. 984–985, **2010**.

<u>Abstract</u>: A study of failures in interconnected networks highlights the vulnerability of

tightly coupled infrastructures and shows the need to consider mutually dependent network properties in designing resilient systems.

[39] D. Balcan, V. Colizza, B. Gonçalves, H. Hu, J. J. Ramasco, and A. Vespignani, “Multiscale mobility networks and the spatial spreading of infectious diseases.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 51, pp. 21484–21489, **2009**.

Abstract: Among the realistic ingredients to be considered in the computational modeling of infectious diseases, human mobility represents a crucial challenge both on the theoretical side and in view of the limited availability of empirical data. To study the interplay between short scale commuting flows and long-range airline traffic in shaping the spatiotemporal pattern of a global epidemic we (i) analyze mobility data from 29 countries around the world and find a gravity model able to provide a global description of commuting patterns up to 300 kms and (ii) integrate in a worldwide-structured metapopulation epidemic model a timescale-separation technique for evaluating the force of infection due to multiscale mobility processes in the disease dynamics. Commuting flows are found, on average, to be one order of magnitude larger than airline flows. However, their introduction into the worldwide model shows that the large-scale pattern of the simulated epidemic exhibits only small variations with respect to the baseline case where only airline traffic is considered. The presence of short-range mobility increases, however, the synchronization of subpopulations in close proximity and affects the epidemic behavior at the periphery of the airline transportation infrastructure. The present approach outlines the possibility for the definition of layered computational approaches where different modeling assumptions and granularities can be used consistently in a unifying multiscale framework.

[40] H. Kim, Z. Toroczkai, P. L. Erdős, I. Miklós, and L. Á. Székely, “Degree-based graph construction,” *J. Phys. A Math. Theor.*, vol. 42, no. 39, p. 12, **2009**.

Abstract: Degree-based graph construction is a ubiquitous problem in network modelling (Newman et al 2006 The Structure and Dynamics of Networks (Princeton Studies in Complexity) (Princeton, NJ: Princeton University Press), Boccaletti et al 2006 Phys. Rep. 424 175), ranging from social sciences to chemical compounds and biochemical reaction networks in the cell. This problem includes existence, enumeration, exhaustive construction and sampling questions with aspects that are still open today. Here we give necessary and sufficient conditions for a sequence of nonnegative integers to be realized as a simple graph’s degree sequence, such that a given (but otherwise arbitrary) set of connections from an arbitrarily given node is avoided. We then use this result to present a swap-free algorithm that builds all simple graphs realizing a given degree sequence. In a wider context, we show that our result provides a greedy construction method to build all the f-factor subgraphs (Tutte 1952 Can. J. Math. 4 314) embedded within $K_n \setminus S_k$, where K_n is the complete graph and S_k is a star graph centred on one of the nodes.

[41] D. Balcan, V. Colizza, A. C. Singer, C. Chouaid, H. Hu, B. Gonçalves, P. Bajardi, C. Poletto, J. J. Ramasco, N. Perra, M. Tizzoni, D. Paolotti, W. Van den Broeck, A.-J. Valleron, and A. Vespignani, "Modeling the critical care demand and antibiotics resources needed during the Fall 2009 wave of influenza A(H1N1) pandemic.," *PLoS Curr.*, vol. 1, p. RRN1133, **2009**.

> Abstract: While the H1N1 pandemic is reaching high levels of influenza activity in the Northern Hemisphere, the attention focuses on the ability of national health systems to respond to the expected massive influx of additional patients. Given the limited capacity of health care providers and hospitals and the limited supplies of antibiotics, it is important to predict the potential demand on critical care to assist planning for the management of resources and plan for additional stockpiling. We develop a disease model that considers the development of influenza-associated complications and incorporate it into a global epidemic model to assess the expected surge in critical care demands due to viral and bacterial pneumonia. Based on the most recent estimates of complication rates, we predict the expected peak number of intensive care unit beds and the stockpile of antibiotic courses needed for the current pandemic wave. The effects of dynamic vaccination campaigns, and of variations of the relative proportion of bacterial co-infection in complications and different length of staying in the intensive care unit are explored.

[42] A. Vespignani, "Predicting the behavior of techno-social systems," *Science* vol. 325, no. July, pp. 425–428, **2009**.

> Abstract: We live in an increasingly interconnected world of techno-social systems, in which infrastructures composed of different technological layers are interoperating within the social component that drives their use and development. Examples are provided by the Internet, the World Wide Web, WiFi communication technologies, and transportation and mobility infrastructures. The multiscale nature and complexity of these networks are crucial features in understanding and managing the networks. The accessibility of new data and the advances in the theory and modeling of complex networks are providing an integrated framework that brings us closer to achieving true predictive power of the behavior of techno-social systems.

[43] F. Schweitzer, G. Fagiolo, D. Sornette, F. Vega-Redondo A. Vespignani, D. R. White "Economic Networks," *Science.*, vol. 325, no. July, pp. 422–426, **2009**.

> Abstract: The current economic crisis illustrates a critical need for new and fundamental understanding of the structure and dynamics of economic networks. Economic systems are increasingly built on interdependencies, implemented through trans-national credit and investment networks, trade relations, or supply chains that have proven difficult to predict and control. We need, therefore, an approach that stresses the systemic complexity of economic networks and that can be used to revise and extend established paradigms in economic theory. This will facilitate the design of policies that reduce

conflicts between individual interests and global efficiency, as well as reduce the risk of global failure by making economic networks more robust.

[44] P. L. Erdős, I. Miklós, and Z. Toroczkai, "A simple Havel-Hakimi type algorithm to realize graphical degree sequences of directed graphs," *El. J. Comb.* vol. 17, p. 11, **2009**.

Abstract: One of the simplest ways to decide whether a given finite sequence of positive integers can arise as the degree sequence of a simple graph is the greedy algorithm of Havel and Hakimi. This note extends their approach to directed graphs. It also studies cases of some simple forbidden edge-sets. Finally, it proves a result which is useful to design an MCMC algorithm to find random realizations of prescribed directed degree sequences.

[45] D. Balcan, H. Hu, B. Goncalves, P. Bajardi, C. Poletto, J. J. Ramasco, D. Paolotti, N. Perra, M. Tizzoni, W. Broeck, V. Colizza, and A. Vespignani, "Seasonal transmission potential and activity peaks of the new influenza A(H1N1): a Monte Carlo likelihood analysis based on human mobility," *BMC Med.*, vol. 7, no. 1, p. 45, **2009**.

Abstract: Background: On 11 June the World Health Organization officially raised the phase of pandemic alert (with regard to the new H1N1 influenza strain) to level 6. As of 19 July, 137,232 cases of the H1N1 influenza strain have been officially confirmed in 142 different countries, and the pandemic unfolding in the Southern hemisphere is now under scrutiny to gain insights about the next winter wave in the Northern hemisphere. A major challenge is pre-empted by the need to estimate the transmission potential of the virus and to assess its dependence on seasonality aspects in order to be able to use numerical models capable of projecting the spatiotemporal pattern of the pandemic. Methods: In the present work, we use a global structured metapopulation model integrating mobility and transportation data worldwide. The model considers data on 3,362 subpopulations in 220 different countries and individual mobility across them. The model generates stochastic realizations of the epidemic evolution worldwide considering 6 billion individuals, from which we can gather information such as prevalence, morbidity, number of secondary cases and number and date of imported cases for each subpopulation, all with a time resolution of 1 day. In order to estimate the transmission potential and the relevant model parameters we used the data on the chronology of the 2009 novel influenza A(H1N1). The method is based on the maximum likelihood analysis of the arrival time distribution generated by the model in 12 countries seeded by Mexico by using 1 million computationally simulated epidemics. An extended chronology including 93 countries worldwide seeded before 18 June was used to ascertain the seasonality effects. Results: We found the best estimate R0 = 1.75 (95% confidence interval (CI) 1.64 to 1.88) for the basic reproductive number. Correlation analysis allows the selection of the most probable seasonal behavior based on the observed pattern, leading to the identification of plausible scenarios for the future unfolding of the pandemic and the estimate of pandemic activity peaks in the different hemispheres. We provide estimates for the number of hospitalizations and the attack rate for the next wave as well as an extensive sensitivity analysis on the disease parameter values. We also studied the effect of systematic therapeutic use of antiviral drugs on the epidemic timeline. Conclusion: The analysis shows the potential for an early epidemic peak occurring in October/November in the Northern hemisphere, likely before large-scale vaccination campaigns could be carried out. The baseline results refer to a worst-case scenario in which additional mitigation policies are not considered. We suggest that the planning of additional mitigation policies such as systematic antiviral treatments might be the key to delay the activity peak in order to restore the effectiveness of the vaccination programs.

IV. Technical highlights

A) Geo-localized WMD (G-WMD) related work

1) A multiscale approach to network centrality measures

Multiscale analysis of complex, real networks requires efficient algorithms that are able to analyze the network structure on all scales, from the local (node) level to the global (graph diametric) level and can extract those subgraphs/structures that are most responsible for the network's principal modes of behavior. While graph algorithms for local level analysis are well developed, there are only few efficient algorithms for a systematic *scan* of the graph scales, for example, for its transport properties. In particular, when trying to determine the vulnerabilities of a network, or the consequences of damages to a network following a WMD attack, one has to be able to provide an accurate description of the damage levels and consequences across the *full range of scales*. There are numerous examples when a relatively small located damage affects network behavior long-distance, causing massive changes in transport function (e.g., cascading failures in power grids). To identify these damage "hot spots" and their consequences over scales we have developed a systematic approach that ultimately provides the network's vulnerability backbone [6], [7], as presented in the proposal. The vulnerability backbone is the smallest structure whose removal or damage causes the largest disruption in the network.

The vulnerability backbone is detected using a betweenness centrality percolation approach, as described in the proposal. Betweenness centrality and its many variants [7], describes the positioning "importance" of a structure of interest such as a node, edge, or subgraph with respect to the whole network. Betweenness centrality lies at the core of both the transport and structural vulnerability properties of complex networks. Betweenness centrality of a node (edge, or a subgraph) is defined as the fraction of all-node-pair (all source-destination pair) shortest paths running through that node (edge or subgraph). Since transport tends to minimize the cost or time of the route from source to destination, it expectedly happens along network geodesics (shortest- or lowest cost paths), and therefore centrality measures are typically defined as a function of these, however generalizations to arbitrary distributions of transport paths have also been introduced. Geodesics are important for structural connectivity as well: removing nodes (edges) with high BC, one obtains a rapid increase in diameter, and eventually the structural breakup of the graph. Nodes (edges) with high BC values play a critical role in network transportation and thus damages to them have the largest consequences on transport performance. Note that once a node or edge is removed, the distribution of the shortest paths (SPs) changes, and accordingly, the BC values also need to be recalculated in order for the

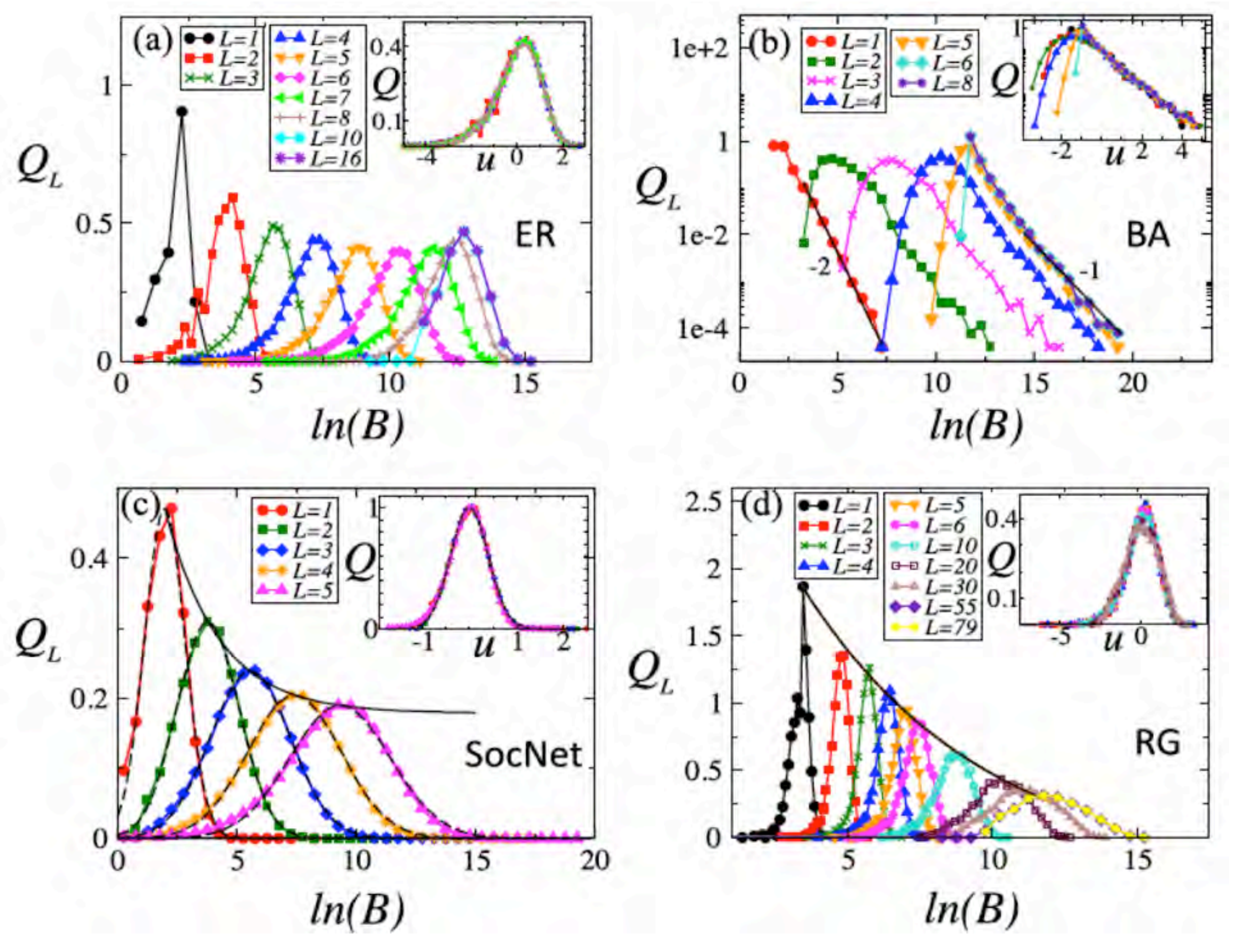


Fig 1 : A universal scaling of range-limited betweenness centrality in large networks. Erdos-Renyi graphs (a), Barabasi-Albert scale-free graphs, a real-world social network inferred from cell-phone communication logs with $N = 5\,568\,785$ nodes and $M = 26\,822\,764$ edges (c) and finally for Random Geometric graphs in the plane (d).

breakup process to be the most efficient. While BC is an important and useful quantity, it is also computationally costly, and its measurement for networks with millions of nodes will be very time consuming. In this project we have introduced a more refined approach to network centrality measures, the so-called *range-limited centrality*, which is able to provide a multiscale view of network centrality and it is also computationally efficient. In particular, range-limited betweenness centrality (LBC) computes the fraction of shortest paths that are passing through a node (or edge) coming from paths not longer than a given value, L. Here L acts as a parameter of the multiscale description. In the limit when L is equal to the diameter D of the graph, we recover the original definition for BC. However, now instead of a single BC value associated with anode or edge we obtain a set (vector) of LBC values, containing lot more information about the positioning importance of the node/edge in the network than a single number. We have developed a corresponding algorithm [6] that generates these LBC values efficiently, for all nodes and edges in a network. We have first formulated the algorithm for shortest-path (SP) betweenness [6] then generalized it to lowest path weight (LPW), or lowest path cost betweenness in [7]. In the latter case the betweenness of a node or edge is the number of *all-pair* LPW paths that run through that node or edge. In general, calculating all-pair paths is the most time consuming part of all centrality type algorithm and it will simply fail on networks with millions of nodes or larger. We need an algorithm that circumvents this to a good extent. In addition, to better assess damages from the *point of view of a node*, or of a network region, with the source of the damage coming from within some given distance on the network (in terms of hop-counts) we need an algorithm that measures betweenness centralities within a given range path length.

The efficient limited-range betweeness algorithm we built [6] can be used for *directed networks as well,* which calculates the SP-betweeness value of a node (edge) for all (directed) paths emanating from that node (edge) of a given length (binary or hop-count) l. Producing these values for all nodes (edges) we obtain a list, from which one can construct betweennesses associated with arbitrary subgraphs. For new l values, the algorithm does not recalculate all the paths, but it uses the previous lists in a smart way to generate the new values, and this considerably speeds up the algorithm.

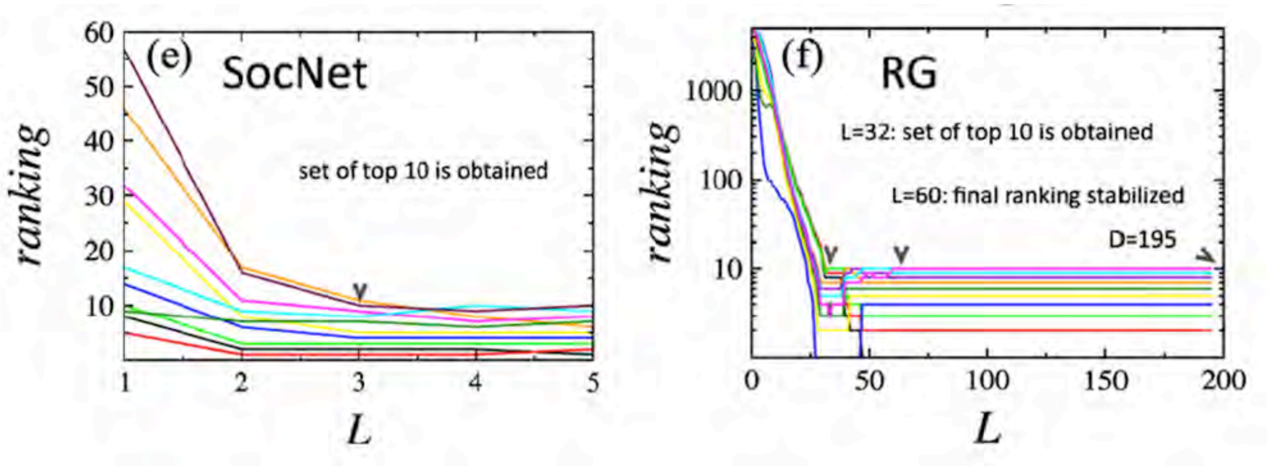


Fig 2 : Fast freezing of node ranks by betweenness: a) for a large (5.5 million nodes, 27 million edges) social network inferred from cell-phone communication logs and for b) Random Geometric graphs (RGG). The plots show the ranking of the 10 nodes with top betweenness values as the range limit is increased. Random geometric graphs are simple models of spatially embedded random graphs. For example, in 2D, an RGG is obtained by first randomly placing points within a domain and connecting any two points with an edge that are less than a given distance apart. RGG-s can be considered as the simplest models of an ad-hoc network of transceivers.

By introducing such a multiscale decomposition of shortest paths, we have shown [6] that the contributions to betweenness coming from geodesics not longer than L obey a *universal scaling* versus L, which can then be used to predict the distribution of the full centralities (by simply extrapolating from the scaling relation). This universal scaling can be interpreted as a *central limit theorem acting on the level of paths*. We have developed a mathematical formalism for this scaling and have shown that the scaling of centrality measures is related to the scaling of the size of the

l-th order shell (the set of nodes at shortest path distance l) around a randomly chosen node. The formula behind this scaling is [6]:

$$\rho_l(b) = \frac{1}{b} \int dk\, P(k)\Phi_l\left(\ln b - \ln \beta_l - \ln k\right) \tag{1}$$

where b denotes node betweenness, $\rho_l(b)$ is the probability density of nodes with l-betweenness value of b. $P(k)$ is the degree distribution in the network, k denotes the degree (approximated as a continuous variable here) and $\beta_l = \frac{l+1}{2}\frac{\langle z_l \rangle}{\langle k \rangle}$. Here $\langle z_l \rangle$ is the average number of nodes found at distance l from a central or root node also called average shell size (average over all nodes as root nodes) and $\langle k \rangle$ is the average degree. In many networks the shell size grows exponentially with l, and in this case $\beta_l \sim \alpha^l$ where $\alpha = \frac{\langle z_2 \rangle}{\langle k \rangle}$. In other networks, such as spatial networks, however, the shell size grows as a power law $\beta_l \sim l^d$ where d is the embedding dimension (for example for roadway networks $d = 2$). Φ_l in (1) is the distribution of noise in the shell size of order l, with $\Phi_1(x) = \delta(x)$ (Dirac-delta). Φ_l was often found to be Gaussian for real-world network datasets. For the class of networks in which the shell size distribution has finite moments, the centrality measures are rescaled to a Gaussian distribution, whereas for those with diverging moments of the shell-size distribution, the corresponding limit distribution is a standard Lévy-stable distribution (scale-free networks). These scaling results and formula (1) are applicable to all large networks. Figure 1 shows this scaling for several, fundamentally different types of networks. We have also shown that due to the existence of this scaling, we don't need to do full, diameter-path length based measurements of centralities (which is unfeasibly costly) but we can *predict centralities* and their distribution

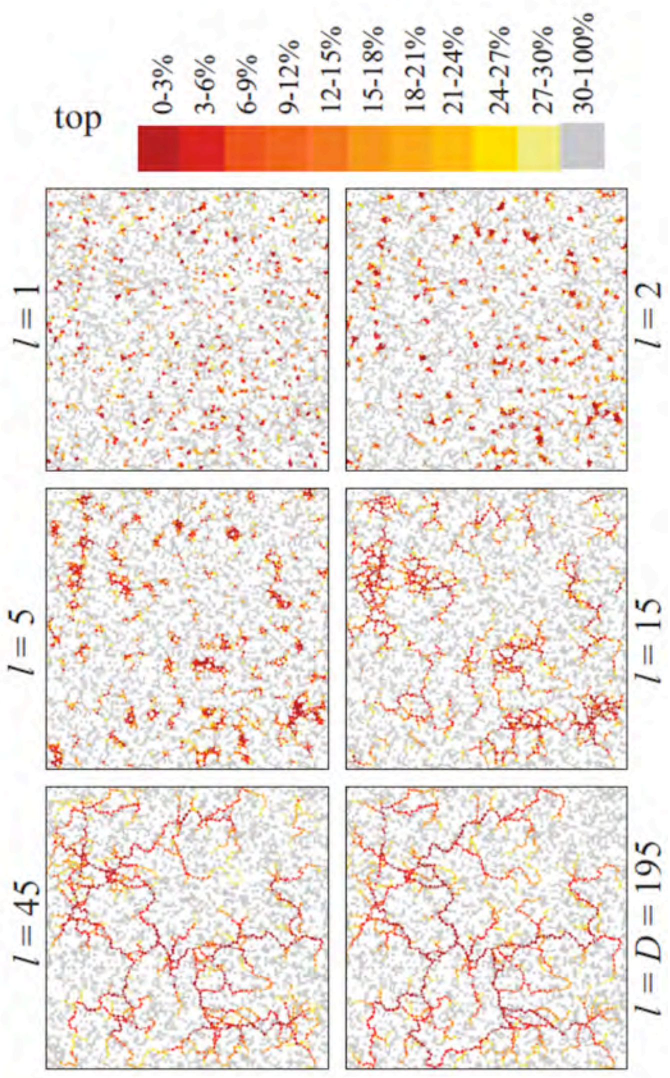


Fig 3 : The vulnerability backbone VB of a random geometric graph in the unit square with $N = 5000$ nodes, average degree $\langle k \rangle = 5$, and diameter$D = 195$. The top 30% of nodes are colored from red to yellow according to their l-BC ranking (see color bar). The VB based on the l-BC is shown for different values: $l = 1,2,5,15,45,195$.

for all path lengths, up to diameter length. One can also use this scaling relation to show that the ranking of nodes and edges by betweenness freezes quickly with the path length L, and hence we can identify *early* the nodes and edges with top betweenness values with low computational cost, as shown in Fig 2. As a matter of fact, while the principal observation (for example, the size of the largest component after subgraph removal based on betweenness values) changes little quantitatively (same qualitative behavior), the computational costs of going from $l = 4$ length paths to $l = 6$ can add a very large computational overhead for only a modest improvement in benefit (increase in costs by several orders of magnitude). With considerably less computational costs, our algorithm has allowed us to identify the vulnerability backbone of a network, which is the collection of highest betweenness nodes and edges that form a percolating subgraph in the whole network. In the context of the cell-phone social network, this backbone is the set of people and relationships that are responsible for the fast spreading of information, the network of super spreaders. Figure 3 shows the convergence of the backbone identified from limited range L betweenness measurements to the final (diameter based structure). This paper has been published in Physical Review Letters, the top technical journal in physics [6]. Next, we first briefly describe the algorithm for un-weighted networks (i.e., for shortest paths), followed by a brief description for the weighted case (lowest weight paths).

Definitions and notations:

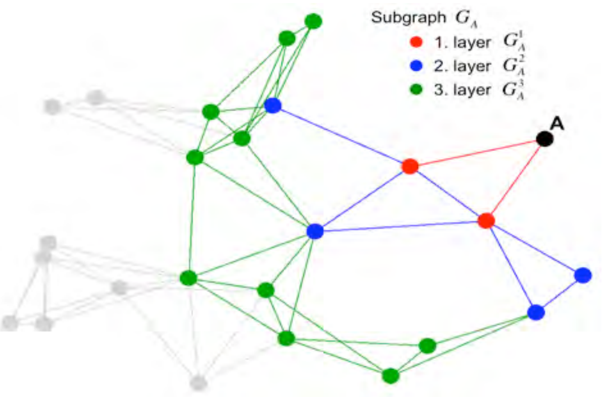


Fig 4 : The subgraph G_A of node A ($= v_i$) . Different layers are shown with different colors: the first layer G_A^1 is shown in red, the second layer G_A^2 in blue, and the third G_A^3 in green.

- $s_{i\to j}$ is the length of the (directed) shortest path between vertices v_i and v_j
- b_i^l is the *l*-betweenness of node v_i for a given length *l*, i.e., the number of shortest (directed) paths running through v_i of exactly length *l*.
- $b_{i\to j}^l$ is the corresponding *l-betweenness* for (directed) edge $e_{i\to j}$.
- $G_i(L)$ is the *L*-range subgraph of node v_i: For a given length *L* it is the subgraph centered on node v_i containing all nodes and edges which can be reached from v_i in no more than *L* steps (along directed paths).
- Layer G_i^l of a subgraph: includes all nodes v_j for which the (directed) shortest path between nodes v_i and v_j is exactly *l*, $s_{i,j} = l$ (see Fig 4).
- $n_{i\to j}$ is the number of possible shortest paths between vertices v_i and v_j .
- $n_{i\to G_i^l}$ is the number of shortest paths between v_j and *all* nodes of layer G_i^l .

The Algorithm for the binary case:

This algorithm generates the *l-betweenness* values of all nodes (b_i^l) and all edges ($b_{i\to j}^l$) in a given graph for all *l =1,2,…,L*.

The algorithm proceeds through all nodes v_i , *i=1,…,N*. For a given node v_i it detects and counts all the shortest paths which start in node v_i and have a length smaller or equal than *L*. Gradually constructing the subgraph $G_i(L)$ it calculates and saves the betweenness values for all nodes and edges involved. The initial node is considered to be layer 0 (G_i^0) and $b_i^0 = 1$. Every node has a sequence of *L* betweenness values, which are being upgraded as new layers are constructed. The following steps are repeated for *l = 1,…,L* :

1. Construct the next layer G_i^l of the subgraph $G_i(L)$ by including all new neighbors of layer G_i^{l-1} .

2. Calculate the *l-betweenness* values b_j^l for all vertices included in this layer $v_j \in G_i^l$, by using the values from the previous layer (Fig 5):

   $$b_j^l = \sum_k b_k^{l-1}$$

   where $v_k \in G_i^{l-1}$, and there is a directed link *k→j.*

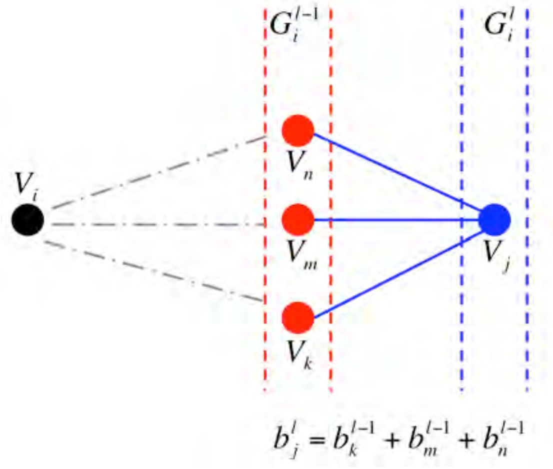


Fig 5: The *l*-betweenness of the new layer *l* is calculated using the *(l-1)*-betweenness values of the nodes in layer *l-1*.

3. Using the *l-betweenness* values of the nodes just included in the new layer the algorithm them gradually calculates the *l-betweenness* values of all edges and nodes of all previous layers (note that these values were all zero before the new layer was added). Thus, for $r = l-1, l-2, \cdots, 1, 0$, the following steps are repeated:

   a. Calculate the *l*-betweenness for all edges between layers *r* and *r + 1*.

   The *l-betweenness* of edge $e_{j\to k}$ ($v_j \in G_i^r$ and $v_k \in G_i^{r+1}$) equals to the product of the number of shortest paths from v_i to v_j, and the number of shortest paths from layer *l* (G_i^l) to node v_k (**Fig 6**):

   $b_{j\to k}^l = n_{i\to j} n_{k\to G_i^l}$.

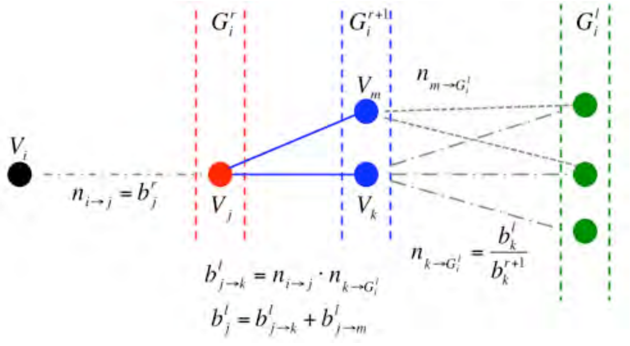

G_i^r
G_i^{r+1}
G_i^l
V_m
$n_{m\to G_i^l}$
V_i
$n_{i\to j} = b_j^r$
V_j
V_k
$n_{k\to G_i^l} = \frac{b_k^l}{b_k^{r+1}}$
$b_{j\to k}^l = n_{i\to j} \cdot n_{k\to G_i^l}$
$b_j^l = b_{j\to k}^l + b_{j\to m}^l$

Fig 6: The *l*-betweenness values of the edges between layers r and $r + 1$ and nodes in layer r are calculated using the previously calculated betweenness values of the nodes in layer $r + 1$.

   Because v_j is in layer *r*, the number of shortest paths from v_i and v_j equals the *r*-betweenness value of v_j, which was previously calculated and saved during the algorithm:

   $n_{i\to j} = b_j^r$.

   The number of shortest paths from

v_k found in layer G_i^{r+1} and layer G_i^l is obtained as the ratio between the *l*-betweenness and the $(r+1)$-betweenness of v_k:

$$n_{k \to G_i^l} = \frac{b_k^l}{b_k^{r+1}} \ .$$

Thus, finally, the l-betweeness of edge $e_{i \to k}$ can be written as:

$$b_{j \to k}^l = b_j^r \frac{b_k^l}{b_k^{r+1}} \ .$$

b. Calculate the l-betweeness of all nodes included in layer r. For a given node $v_j \in G_i^r$ this can be calculated by adding the betweenness values of all edges between v_j and layer G_i^{r+1}

$$b_j^l = \sum_k b_{j \to k}^l \ ,$$

where there is a directed link from v_i to v_k and $v_k \in G_i^{r+1}$ (Fig 6).

4. When the *l*-betweenness values of all layers were calculated, we return to step 1 and find a new layer, until all *L* layers are constructed.

This algorithm calculates the betweenness values from all shortest paths (of maximum length L) starting from vertex v_i . To find the total l-betweenness for every node (edge), we have to repeat this algorithm for every node v_i , $i = 1, \cdots, N$ (and edge) and sum the corresponding values.

The Algorithm for the weighted case

Calculating betweenness centrality of a node or edge in a directed graph $G(V,E)$ requires to count the number of all-pair shortest directed paths incident on it. Betweenness centrality (BC) is defined as:

$$B(i) = \sum_{m,n \in V} \frac{\sigma_{mn(i)}}{\sigma_{mn}} \qquad (2)$$

where σ_{mn} is the number of all possible shortest paths from node m to node n and $\sigma_{mn}(i)$ is the number of shortest paths from node m to node n passing through node i. Similar quantities can be defined for an edge $(i,j) \in E$: $B(j,k) = \sum_{m,n \in V} \sigma_{mn}(j,k)/\sigma_{mn}$.

In un-weighted graphs the length of the shortest path between two nodes is measured as the number of edges included in the shortest path, or the smallest number of hops along edges needed to reach one node starting from the other. In weighted networks each edge has a weight or length: w_{ij}. Depending on the nature of the network, this length can be an actual physical distance value (for example in road networks), or any resistance-type quantity. The shortest path between nodes i and j is defined as the path along which the sum of the length of the edges included is minimal. We will call this sum as the distance $d(i,j)$ the two nodes.

In order to define a range-limited quantity, let $b_l(j)$denote the betweenness centralities of a node *j* for all-pair lowest cost directed paths of length $D_{l-1} < d \leq D_l$, where $D_1 < D_2 < \ldots < D_L$ are a series of predefined distances, which we will call regions. The simplest way to define these D_l distances is to take them uniformly as $D_l = l\ \Delta d$, however, depending on the application, these can be defined in other ways as well. We will denote the cumulative L-betweenness as $B_L(j) = \sum_{l=1}^{L} b_l(j)$, which represents centralities from paths not longer than D_L. Similar measures for an edge are defined in the same way. Just as all centrality algorithms, our method calculates these quantities for a node *j* for shortest directed paths all emanating from a "root" node *i*, then it sums the obtained values for all $i \in V$ to get the final centralities for *j* (similarly for edges). We will derive recursions that simultaneously compute the BC for all nodes and edges and for all regions $l = 1, \cdots, L$. The algorithm's output thus generates detailed and systematic information about shortest paths in a weighted graph on all length-scales, providing a tool for multiscale network analysis.

The algorithm starts from a given root *i* and builds the L-range subgraph C_L containing all nodes which are closer than D_L. Only links which are part of the shortest paths starting from the root are included in C_L. We decompose C_L into shells $G_l(i)$ containing all the nodes at shortest path distance $D_{l-1} < d \leq D_l$ from the root. The root itself is considered to be shell *0* ($G_0(i)$). Notice that in this case there can be edges connecting nodes which are not in two consecutive layers, but possibly further from each other, or occasionally, even in the same layer. An edge *j->k* is considered to be part of the layer in which node *k* is included.

When building the subgraph, we will need to save the exact order in which the nodes and edges are included in the subgraph. Let us denote with *z(j)* the index of the node which is included at the j^{th} place in this list. This will mean that for the distances of the nodes $z(1), z(2)$, etc. the following conditions hold: $d(i,z(1)) \leq d(i,z(2)) \leq d(i,z(3)) \leq \ldots$. Similarly, we have a list of edges, where $u_x(j)$-> $u_y(j)$ is the edge in the j^{th} place, u_x and u_y denoting the index of the two nodes connected by the edge. This implies the condition that: $d(i,u_y(1)) \leq d(i,u_y(2)) \leq d(i,u_y(3)) \leq \ldots$ When calculating the *l*-betweenness value of a node *j*, which is in the r^{th} shell $G_r(i)$ of node *i*, we will denote this by $b_l^r(i|j)$, similarly for a directed edge $b_l^r(i|j \to k)$. Note that these values take into account only the lowest cost path starting from node *i*.

Our algorithm has the following main steps:

1) We build the next layer $G_l(i)$ using breadth first search. We include in this layer all nodes for which the distance from the root node is larger than D_{l-1} but smaller than D_l. During the breadth-first search we build the list of indices *z*, u_x, u_y as defined above. We also update the σ_{ij} values for all nodes (j) included and the *l*-betweenness is set to $b_l^r(i|j) = 1$;

2) Going backwards through the list of edges (saved in u_x and u_y), we perform the following recursions for each edge $u_x(m) \rightarrow u_y(m)$:

$$b_l^r\left(i|u_x(m) \to u_y(m)\right) = b_l^r\left(i|u_y(m)\right)\frac{\sigma_{iu_x(m)}}{\sigma_{iu_y(m)}} \tag{3}$$

At the same time, the betweenness of node $u_x(m)$ is also updated:

$$b_l^r\left(i|u_x(m)\right) = b_l^r\left(i|u_x(m)\right) + b_l^r\left(i|u_x(m) \to u_y(m)\right) \tag{4}$$

We perform this routine for every root node *i* in the graph. At the end of the algorithm we have the *l*-betweenness values of every nodes and edges for all regions $l = 1, \cdots, L$.

2) Predicting traffic flows in large transportation networks: the case of US roadways

One of the challenges in network science is predicting network flows from graph structural properties, node/ edge attributes and dynamical rules. While for some networks (for example, electronic circuits) this is a well-understood problem, it is still open in general, and especially for networks involving a social component such as communication networks, epidemic networks and infrastructure networks. We have focused on the traffic flow prediction problem in spatial networks, and in particular in roadway networks, and validated our results using US highway network and traffic data (http://libguides.mit.edu/gis). Understanding flows in spatial networks driven by human mobility would have many important consequences: it would enable us to connect throughput properties with demographic factors and network structure; it would inform urban planning; help forecast the spatio-temporal evolution of epidemic patterns, help assess network vulnerabilities and allow the prediction of changes in the wake of catastrophic events such as G-WMD events.

As our goal was to provide a validated, practically usable method to network flow modeling in the wake of single and multiple WMD events and scenario testing, we need to first ensure that our approach is able to reproduce *actual traffic data*; even before WMD event scenarios are introduced. Our results, that we describe briefly below have been published in *Nature Communications* [2], with an additional paper to be submitted.

When modelling transportation systems as networks, we associate network nodes with locations and edges with physical paths between locations. We define nodes as intersections between the roads and the road segment between two consecutive intersections as the edge connecting those nodes. We refer to nodes also as sites or locations, interchangeably. Our ultimate goal is to determine the average traffic flow T_{ij} expressing the number of flow units (for example vehicles) per unit time (for example per day) through an edge (i,j) of the network, given the network and the distribution of the population.

The principle idea behind our flow modeling is that traffic flows can be computed by modified betweenness centrality measures for weighted networks. The weights represent edge costs such as physical distance, travel times and capacity limitations (e.g., number of lanes). Node-to-node transport happens along total-cost minimizing paths. Collating the minimal-cost paths from all node-pairs (a, b) that run through an edge we obtain the set of paths contributing to the average traffic on that edge. However, to compute this traffic, one needs to solve two problems: I. the *Traffic Law Problem*, and II. the *Flux Distribution Problem*.

I. The Traffic Law Problem

is a *socio-demographic* problem. Traffic in the first place is caused by people traveling between sites and thus it depends on the distribution of the population $\rho(\varphi, \lambda)$, where φ and λ are the latitude and longitude. The Traffic Law problem is to find a generative model between fluxes for all source-destination pairs:

$$\{\Phi_{ab}\}_{a,b\in V} = \{\Phi_{ab}(m_a, n_b, r_{ab}, \boldsymbol{K})\}_{a,b\in V}$$

where m_a and n_b are populations at sites a and b, respectively, r_{ab} is the physical distance between the sites (as crow flies) and $\boldsymbol{K}$ is a vector that includes demographic variables such as the number of available jobs. In the second year we have run our simulations with two traffic laws: the so called "gravity law":

$$\Phi_{ab} = c\frac{m_a n_b}{(r_{ab})^{\sigma}}, \qquad \sigma \cong 2$$

and the "exponential gravity law" introduced earlier (also) within this grant:

$$\Phi_{ab} = c\frac{(m_a)^{\alpha}(n_b)^{\beta}}{e^{r_{ab}/\kappa}}.$$

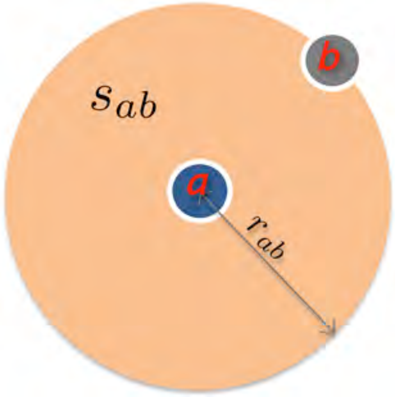


Fig 7: The radiation law uses population density instead of physical distances.

Both models contain adjustable/fitting parameters. However, as we do not have an understanding on how these fitting parameters should be changed once the network has changed (e.g., due to WMD), they are not applicable directly to our case. Instead, we have implemented and further developed a novel traffic law, called the *radiation law*, introduced by Simini et al. [1]. The main advantage of this traffic law is that it has no fitting parameters, except for a single overall scaling constant c corresponding to the average fraction of commuters. The radiation law replaces the direct dependence on distance with population size s_{ab} occupying an area within a ring between the two sites (cities), see Fig 7. This is plausible, as the chief reason for people commuting is the availability of jobs, which correlates with population density. However, people will travel only as far as they find the first suitable job for them. According to this traffic law (see [1] for derivation):

$$\Phi_{ab} = c\frac{{m_a}^2\, n_b}{(m_a + s_{ab})(m_a + n_b + s_{ab})}. \tag{5}$$

II. The Flux Distribution Problem

is a network flow computation problem. According to our method, the annual average daily traffic T_{ij} through an edge (corresponding to a road segment between two intersections) is obtained from traffic fluxes between all those source-destination pairs (a, b) for which edge (i, j) is contained on the minimal cost network path from a to b, see Fig 8.

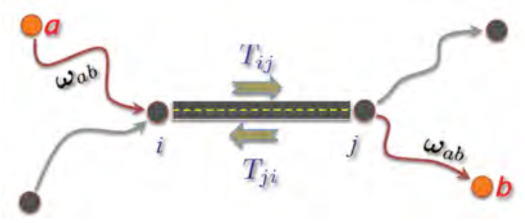


Fig 8: Flows are computed based on a weighted betweenness centrality type measure, from all-pair lowest-cost paths.

$$T_{ij} = \sum_{a,b|(i,j)\in\boldsymbol{\omega}_{ab}} \Phi_{ab} \tag{6}$$

where $\boldsymbol{\omega}_{ab}$ denotes the minimum-cost network path from a to b.This is essentially a betweenness centrality type calculation with weighted edges. However, this computation is rather costly and cannot usually be done online for large networks. This is because we have to take into account $N(N-1) \approx 1.9\times10^{10}$, Φ_{ab} fluxes for a single whole-network betweenness calculation, a calculation with a complexity on the order of $\mathcal{O}(NM) \approx 2.4\times10^{10}$ steps for the US Highway network data (described below). This is unfeasible for WMD scenario analysis (iterated removal and recalculation for various sites and their neighborhoods) for the US Transportation Highway network. In order to resolve this problem, we have expanded our range-limited betweenness calculation algorithm for non-weighted networks to include weighted graphs as well as described in the previous section. This has allowed us to approximate betweenness values much more efficiently. The details of this type of betweenness calculations have been published in Physical Review E see [7]. Additionally, in collaboration with some of our computer science colleagues at Notre Dame we have developed a highly optimized version of this algorithm using a novel, purely pointer-based sparse-row implementation with optimal cache behavior achieving several orders of magnitude speed-up even on large networks containing millions of nodes and tens of millions of edges, as shown in **Fig 9**.

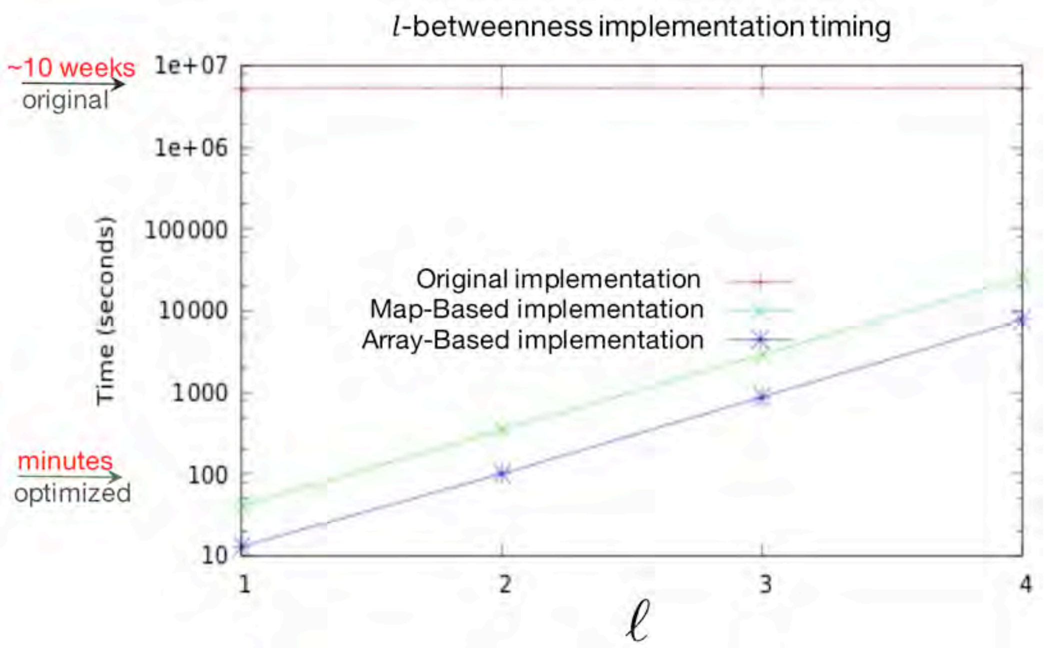


Fig 9: Orders of magnitude speed-up obtained for range-limited l-betweenness calculations using our pointer-based sparse row implementation. The test network used for this simulation was a social network inferred from cell-phone communications having 5 Million nodes and about 27 Million edges.

Data. To validate our flow computing method, we have used US road-network and traffic data from MIT's ArcGIS database. In addition, the population data was obtained from government websites publishing census data at the level of zip codes for the whole contiguous US. This network has $N = 1137\,267$ nodes (highway intersections) and $M = 174\,753$ edges (a network edge is a road segment between two consecutive intersections).

The population data is on a different level (zip-codes) than the network data (nodes as intersections). One problem was how to assign the population to intersections. This was done using a Voronoi tessellation first based on zip-codes, and then distributing the population of a zip code uniformly amongst the intersections within that zip-code, see Fig 10b. The next problem was defining path-cost. First we used physical distances, however, that did not work well. We obtained the best results using minimum *travel time* paths. To compute minimum travel time paths, we used the speed limits associated with the road segments and the physical lengths of the roads. The biggest obstacle was dealing with missing data (missing info on the number of lanes and road class for some of the roads). We used a supervised machine learning approach to predict the missing data based on the existing ones. We validated our predictions on existing data that was not part of the training set to obtain reliable estimates, see [2].

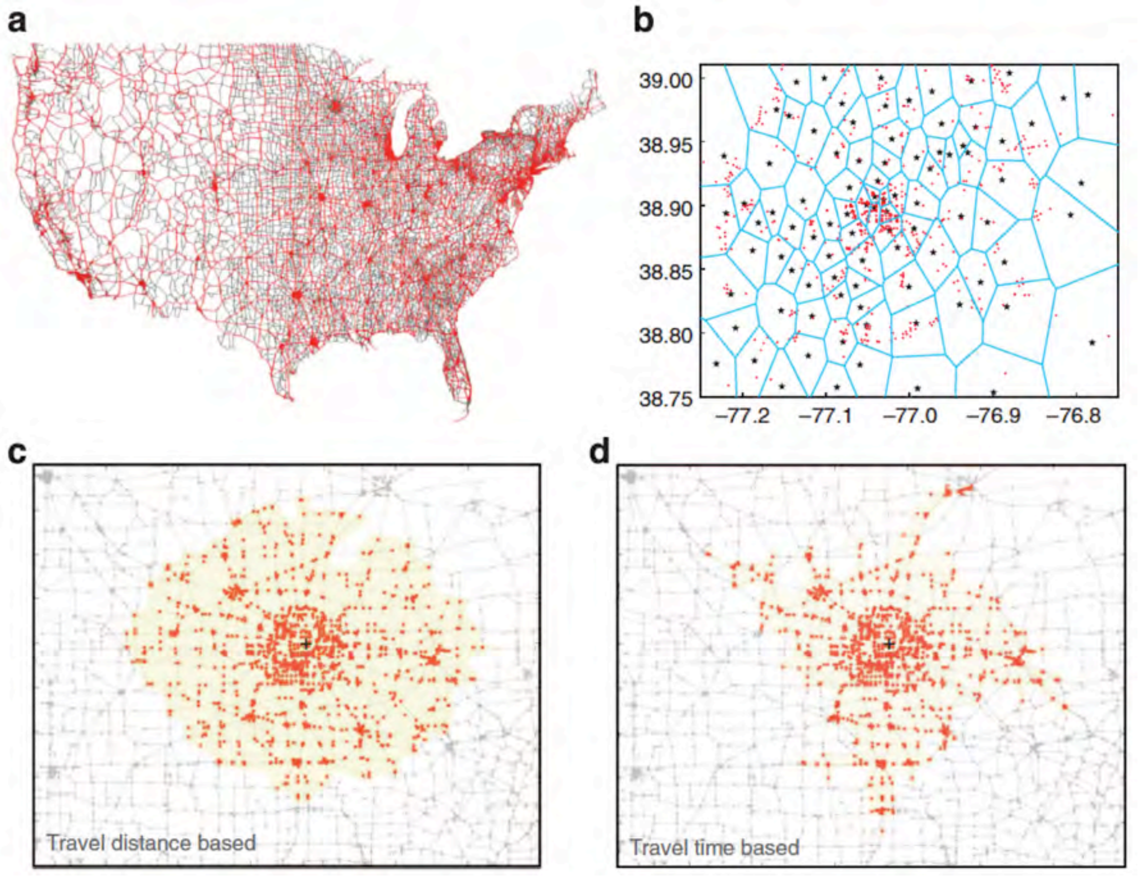


Fig 10: Network and population data. (a) The US highway network with nodes as intersections and edges as road segments between intersections. It has N=137,267 nodes and M=174,753 edges. The red segments (43%) have recorded annual average daily traffic values. (b) Assigning a population size to every intersection (red dots) using a Voronoi mesh and zip-code level census data (zip-code centers indicated by black stars); Washington DC area is shown. (c) Geographical area of locations around a node centered in Minneapolis, MN, with travel cost c_{ab} not larger than a given value using travel distance cab as travel cost. (d) Same as (c), but using travel time τ_{ab} as travel cost.

Even with time-based costs, the original radiation model, although gave better results than any of the gravity models, it was not generating a good Pearson correlation coefficient between traffic values from the data and from the model (on the same roads). The model did, however, capture fairly well the distribution of the fluxes themselves. We eventually realized what was the issue: in our modeling the traffic law problem and the flux distribution problem were decoupled. Indeed, traditional models (including our first attempts) are decoupling these two fundamental problems in the sense that the Traffic Law is solved first and treated independently of the underlying network, and then the fluxes are distributed (2nd problem) on the network using a kind of shortest-path based algorithm.

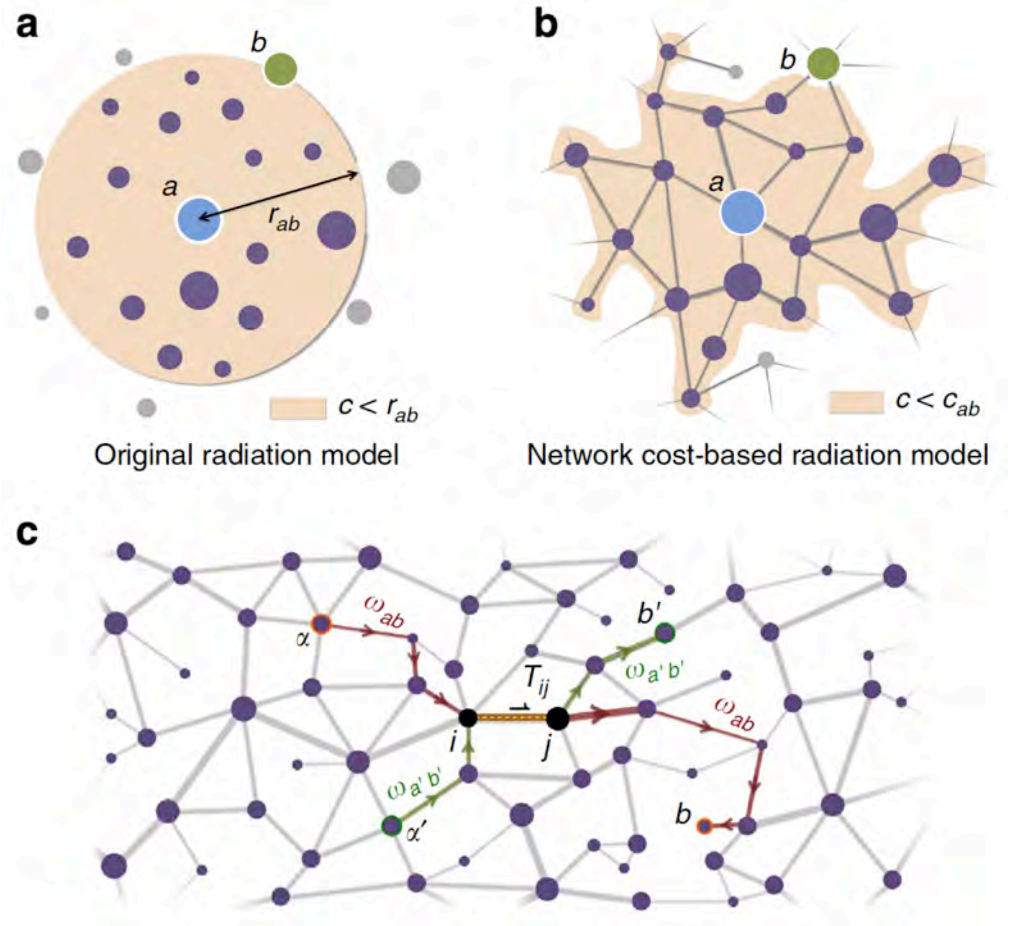


Fig 11: Schematics for traffic flow modelling. (a) The original radiation model uses distance r_{ab} as a search criterion. (b) The cost-based radiation model uses network travel cost c_{ab} as a search criterion, which usually has a heterogeneous distribution (beige shaded region, s_{ab}). (c) The flow T_{ij} through edge (i, j) is the sum of contributions from all those mobility fluxes Φ_{ab} whose minimal cost paths ω_{ab} contain (i, j).

However, we have shown [2] that these two fundamental problems cannot be treated separately if one wants to realistically predict network flows. We have shown that the radiation law giving the travel fluxes between two sites has to be modified such as to take into account the network and the *cost of travel* on the network, see Fig 11.

We have introduced a novel model, the *Extended Radiation Model* that now includes a general cost function c_{ab} (Fig 11) allowing the analyst to compare predictions based on arbitrary cost criteria for travel, including those based on travel distance and travel time. Using a large US traffic dataset from MIT-s database for validation, we have demonstrated that the travel time based cost predictions achieve a significantly better agreement with real traffic than those based on travel

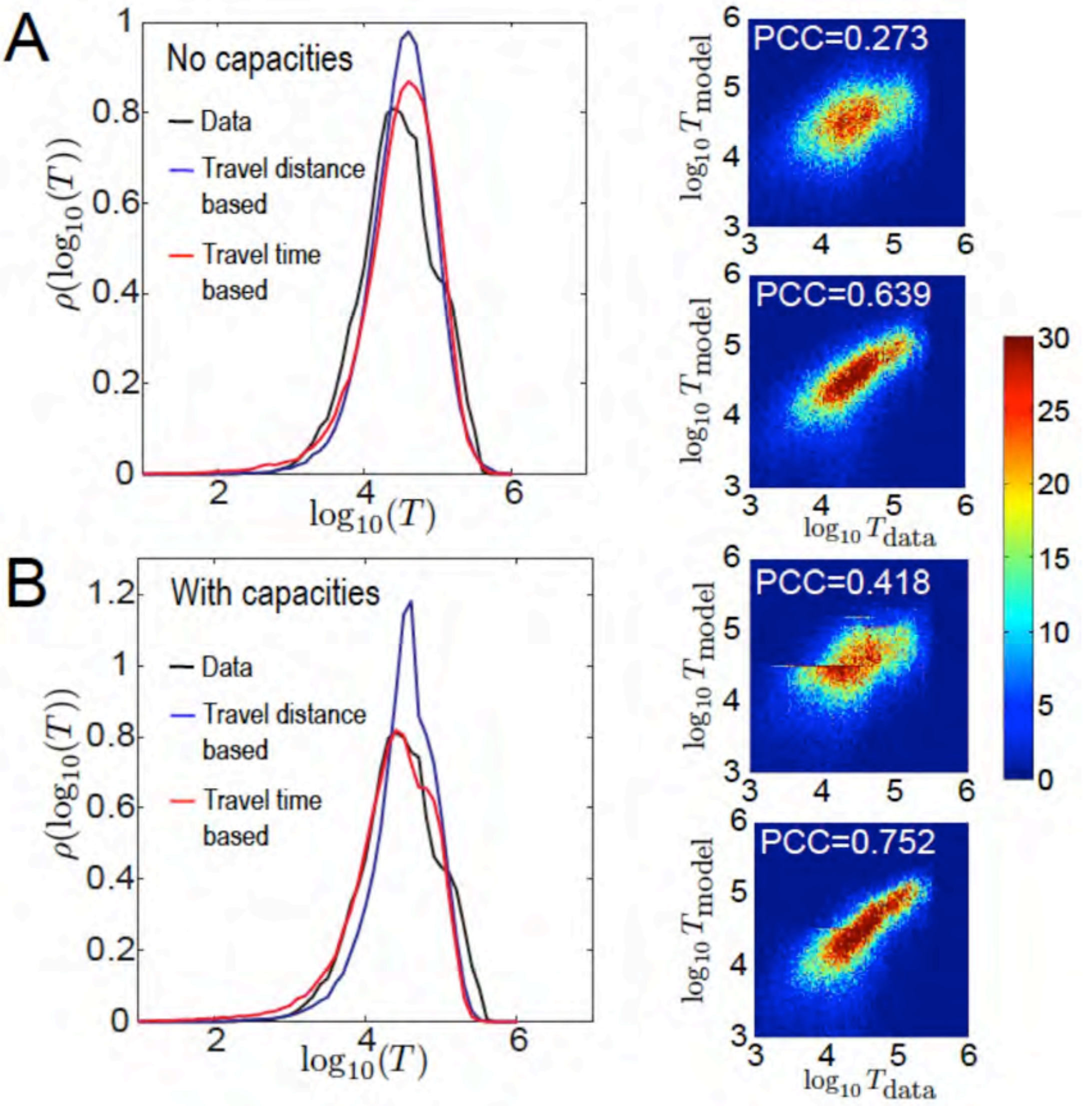


Fig 12: (A) Left panel: comparison between the densities of log(traffic) obtained from data (black line) and the model without capacity limitation based on travel distance (blue line) and travel time (red line). The heat maps are scatter plots between real traffic and model traffic values without capacity limitation. For the upper map the travel distance cost function with a range limit of 100km was used generating a PCC of 0.273. For the lower map the travel time cost function was used with a range limit of 100min and velocity classes 90-40-15mph, generating a PCC of 0.639. (B) Same as in (A) but with capacity limitation included. Here the best PCC of 0.752 was obtained configuration as in (A). The range limits here were 400km and 400min, respectively.

distance alone, see Fig 12. Instead of just comparing distributions of traffic values, we also measure the linear correlation coefficient, or Pearson Correlation Coefficient (PCC) between the model prediction and real data values. The later is a much more stringent test for the goodness of the model than just comparing distributions. A PCC = 1 means perfect agreement. As seen from Fig 12, we could achieve a PCC = 0.752 with our model, which is so far the highest achieved that we know of. While both the travel distance and travel time based cost-functions do well on matching the distribution of flows in the data, only the cost function based on travel times achieves a high linear correlation coefficient (PCC=0.752) between the data and model. This finding suggests that people preponderantly choose travel paths based on a temporal criterion rather than a distance based one. From our analysis it also becomes clear that social behavior is a strong determinant of flows in socio-technical networks and for a realistic modeling of the consequences of a WMD event, we cannot neglect this component.

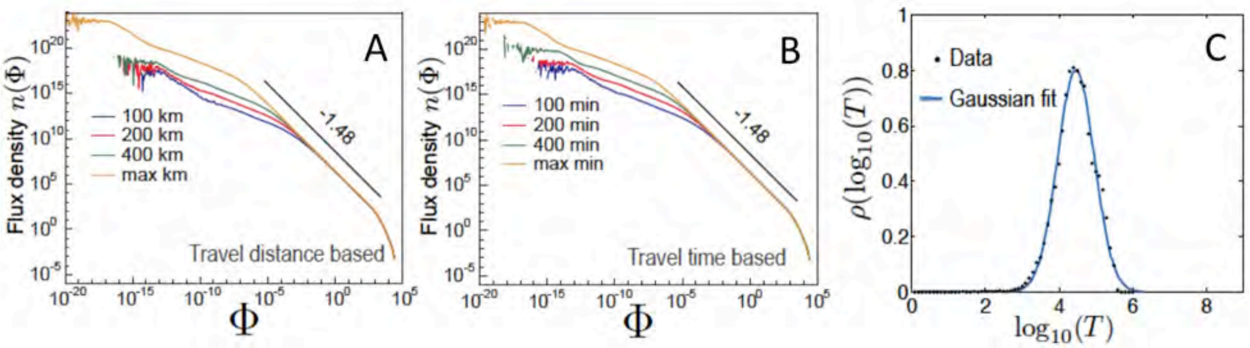


Fig 13: (A,B) Density of source-destination pairs with transport flux Φ based on distance cost function (A) and time based cost function (B). (C) Shows that the distribution of flow values in the network is lognormal (that is, the log of the flows is normally distributed).

We have also made the unexpected discovery that the extended radiation model generates a power-law scaling for the flux density. This scaling holds for over 7 orders of magnitude! See Figs 13A,B. We have then derived this scaling mathematically from purely network-based arguments. The fact that the fluxes become very small beyond a finite range (e.g., already at 100min or 100km in case of traffic flows) it means that range limited computations can be used effectively for flow modeling [7]. Exploiting the direct connection of network flows with betweenness centrality measures and the scaling laws they obey we have also provided mathematical arguments to why the distribution of traffic is a log-normal (see Fig 13C). Figure 14 shows the level of agreement at the specific roadway level between real data and our model generated traffic.

There could exist gravity model versions in the literature that may be used to better match the local flow, but they come at the expense of many additional fitting parameters. However, if we would need to predict new flow patterns in the wake of network changes such as in the wake of

WMD events, it is not clear what values should be used for those fitting parameters *on the changed network*. The main strength of our model is that it is based on *first principles* and thus it can be easily used for flow predictions in the wake of network changes. Our model can be further improved by adding more features, such as actual average speeds on individual roads, a better approximation to population distribution at the intersection level, etc.

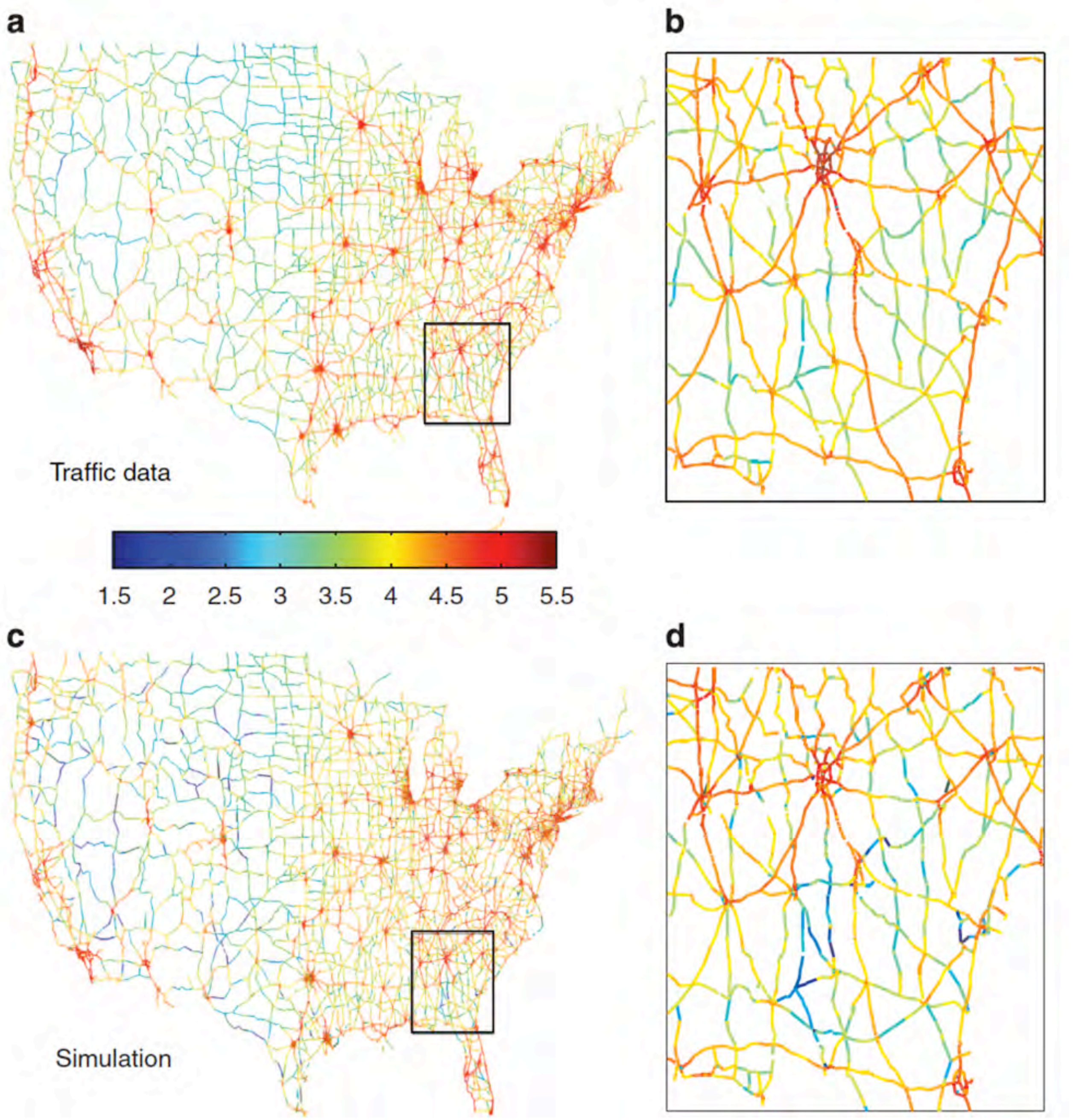


Fig 14: A visual comparison. (a) Log traffic values indicated via colours (see colour bar) on major highways in the contiguous US. (b) Magnification of a south-east region. (c) Same as in (a) but for the model output using travel time cost with capacity limitation and with the same parameters as in Fig. 12b. (d) magnification of the same region from (c) as in (a).

3) Studies of changes in transportation flows as a result of G-WMD events

Given our validated flow modeling methodology we started studying the long-range effects of geo-localized damages on infrastructure networks, and in particular using the validated model for large-scale network flows obtained earlier. As an analysis test bed for our theory, we have used the same large-scale dataset as in the previous case. The geo-localized WMD events we have studied were of two sizes: they correspond to the removal of the portion of the network in a geographical area of 10km radius in one case, and 20km-s in the other case.

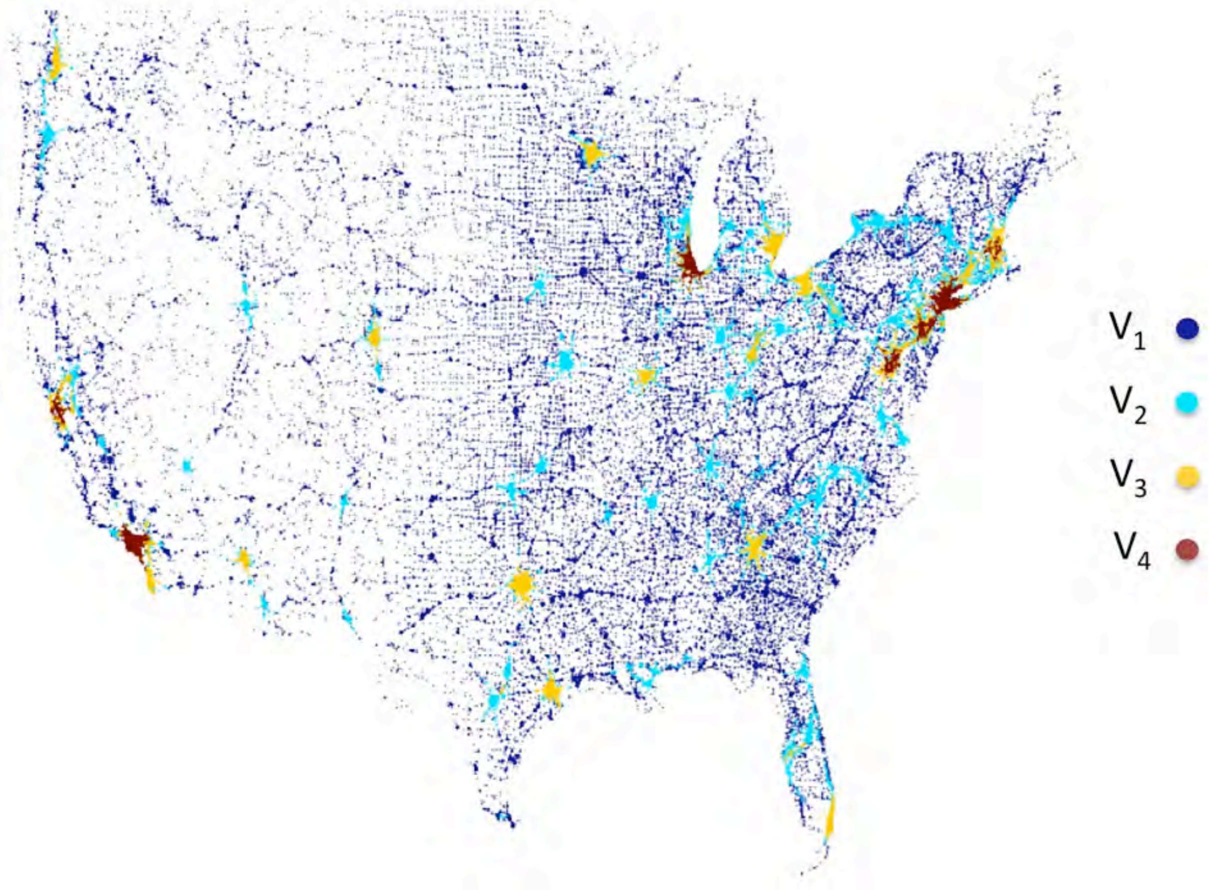
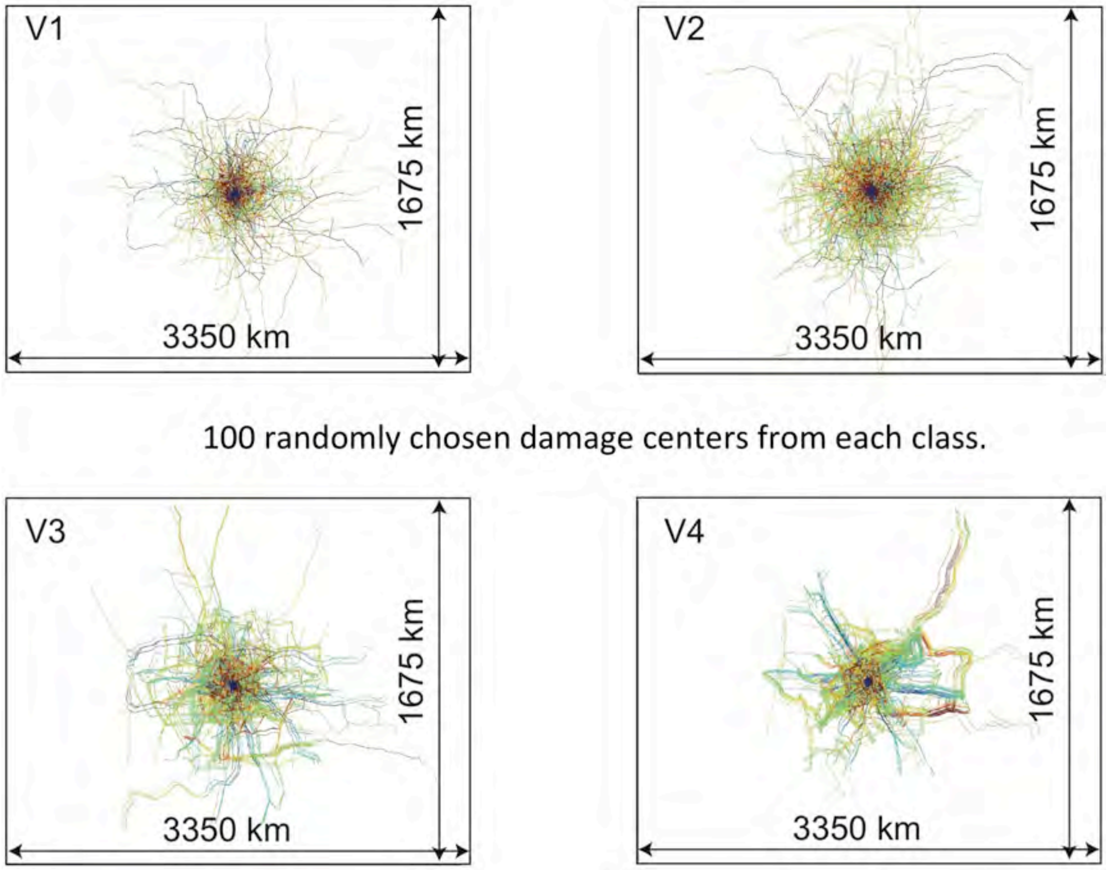


Fig 15: The four classes of nodes used to as ground zero sites for geo-localized WMD events.

The 10km radius case corresponds roughly to the equivalent of a B61 tactical/strategic nuclear weapon's blast at about 340kT explosive yield, whereas the 20km radius corresponds to a B53 strategic nuclear weapon's blast at 9MT yield. Once the part of the network was removed, using our first-principles based flow model we recomputed the flows in the new network and recorded the changes and their distribution. In order to better classify the effects of damages in terms of the location of the WMD event, we have classified the locations according to the sizes of the population that can be reached from the edges of the blast within a given a cost (55 mins time cost). This yielded 4 classes of initial conditions corresponding to population sizes that are roughly equally separated on log-scale, see Fig 15 for their distribution within the continental US. Accordingly, from nodes that are in class V1 we can typically reach a population of 0.2x10^6 people, from class V2 the population reached is 1.24x10^6, from V3 it is 3.32x10^6 and from V4 it is 9.97x10^6. In the continental US there are 69 104 V1 nodes, 32 004 V2 nodes, 20 252 V3 nodes and 15 907 V4 nodes. The latter correspond to large metropolitan areas. The reason that the number of V4 nodes is this large is because we resolve the nodes at the level of road intersections, and thus all the road intersections in large metropolitan areas enter this set. Figure 3 shows the locations of the 4 node types in the continental US. We have selected at random 100 ground zero locations from every node class and overlaid the changes in the network flows in the same coordinate system. Fig 16 shows the magnitude and extent of the damages into the network for the 20km radius affected zone case. Significant flow changes in the network are within a neighborhood of about 300kms, but moderate to weak changes along certain paths can spread to 1000kms or even further.

Both Figs 16, 17 show that the effects have a slow decay with distance. This means



100 randomly chosen damage centers from each class.

Fig 16: Overlaid network flow-change maps from 100 randomly chosen ground zero sites for 20km radius WMD events. Note that the largest non-local effects are when the location of ground zero is within class V2 sites. In case of large metropolitan areas there is massive damage but it is more localized as around such areas there is also an overabundance of network paths.

mathematically, that the effects measured as the change in flow per edge within a ring of a given radius (and thickness of 6km serving as a distance bin) decay slowly, that is algebraically with distance (Fig 17). The long-range effect here is the key observation. This is quite surprising and counter intuitive, in the light that this infrastructure network is a spatial network, and unlike for example an airline transportation network or an information network, it does not have shortcuts or hubs that connect most of the network together and thus would be responsible for a fast spread of damage consequences. In the next section, we perform an in-depth analysis of these observations of long-range effects from geo-localized damages.

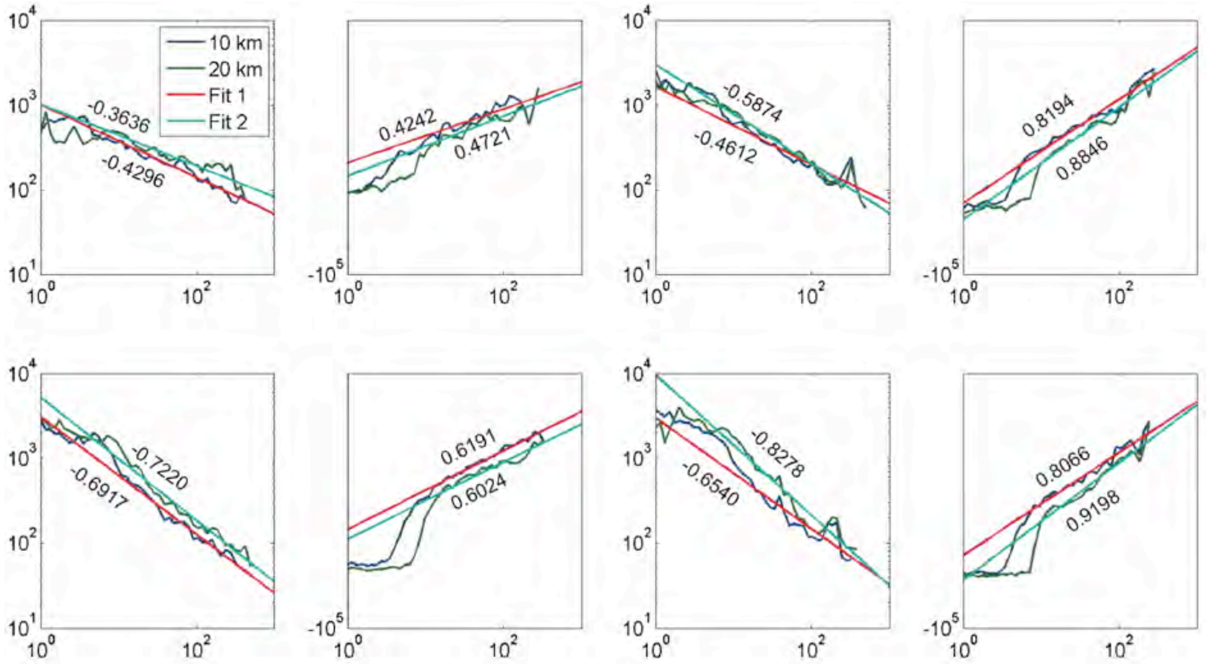


Fig 17: Average change in flow per edge at a given distance (x-axis). Left upper two panels are for V1 sites, the right upper two panels are for V2 sites, left lower two panels are for V3 sites and the right lower two panels are for V4 sites as ground zero sites. The fits are power-laws with small exponents showing that the changes propagate long-range. Within every class, the left panel shows the change for the edges that experience an increase in flow, whereas the right panel corresponds to decrease in flow.

4) Long range and long-term effects from geo-localized WMD events

Previously we have shown the existence of long-range effects from localized damages in infrastructure networks, with applications to WMD. Continuing this line of inquiry, we have performed extended simulations of localized network damage scenarios in the US highway network and studied its impact on transportation flows. Using US census and highway traffic data we were able to show that in spite the fact that roadway networks as spatial networks do not possess shortcuts that would guarantee long-range effects, localized damages can lead to significant traffic changes far away from the damage location. These long-range effects are most significant when the damage center is located in a few, special points in the network, which we call *fragility nodes*. We have shown that the mechanism behind these non-local effects is not based on cascade-propagation, but instead, it is the result of interplay between the heterogeneity in the spatial distribution of the population and the spatial heterogeneity of available network paths.

Figure 18 illustrates two concrete instances of long-range effects: in case a) a circular region of a radius of 30km was removed centered in San Antonio TX and in (b) the same size circular area was removed centered around La Porte, IN. San Antonio is a major city with a population of about 1.4Mill; it is no surprise that it can cause major effects in traffic flows, even far away from the epicenter. However, when looking at the traffic changes after removing the same size circular region centered on La Porte, a small, rural town in Indiana, with a population of only 22K, we not only see major changes, but we observe significant changes far away, around Detroit and Columbus. Epicenters that have such major effects but are otherwise of low population, we call

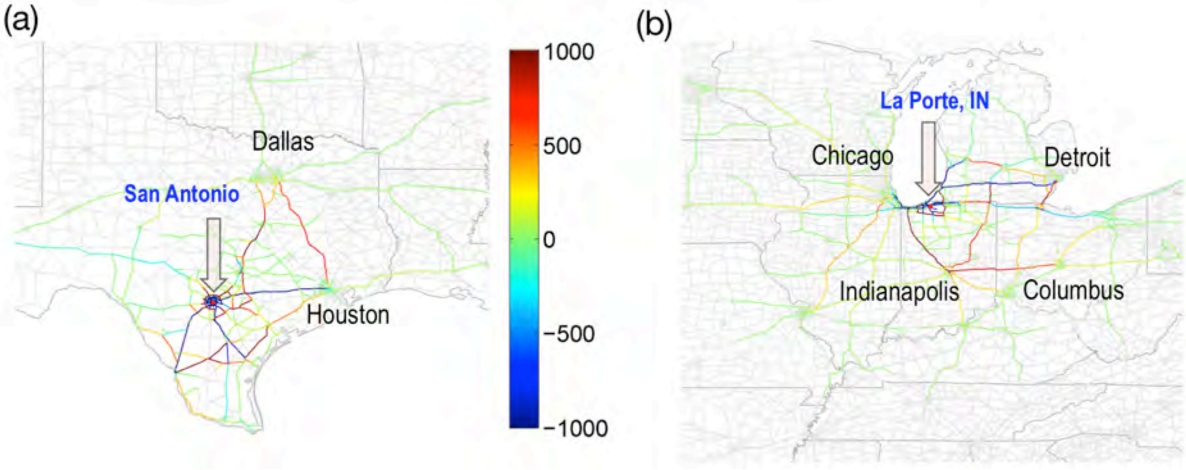


Fig 18: Examples of long-range effects on transportation flow in the wake of a geo-localized WMD event. (a) An event in San Antonio, TX causes long-term traffic increase between Houston and Dallas and other regions (b) An event in La Porte, IN, a very small town (compared to San Antonio) causes major changes far away, including between Indianapolis and Columbus and around Detroit. La Porte is a fragility point example. The color bar shows the daily traffic changes in the number of cars.

network fragility points. Damages at the locations of fragility points have long-range effects on network transportation. In the last period we have been working on understanding the mathematical conditions that generate such network fragility points. The typical scenario after geo-localized network damage in small-population areas is that changes are mostly local, affecting the traffic only in the immediate neighborhood of the damaged area. However, there are a few special regions, like La Porte, that are fragility points with major long-range effects. Figure 2 shows an overlay of traffic flow changes from 1000 damage centers uniformly distributed within US, illustrating that effects can propagate within a radius of 600 miles. What causes these long-range effects and how can we characterize the corresponding fragility points? To answer this question we have performed another set of extended simulations in which we only remove the population from the same damage area, but leave the network intact, meaning that transportation paths that were previously using those paths are being used (by the population in the rest of the network) even after the damage. We have defined the quantities:

$$S_{B(P)} = \sum_{ij}^{B(P)} |\Delta T_{ij}|, \quad \sigma = \frac{S_B}{S_P} \tag{7}$$

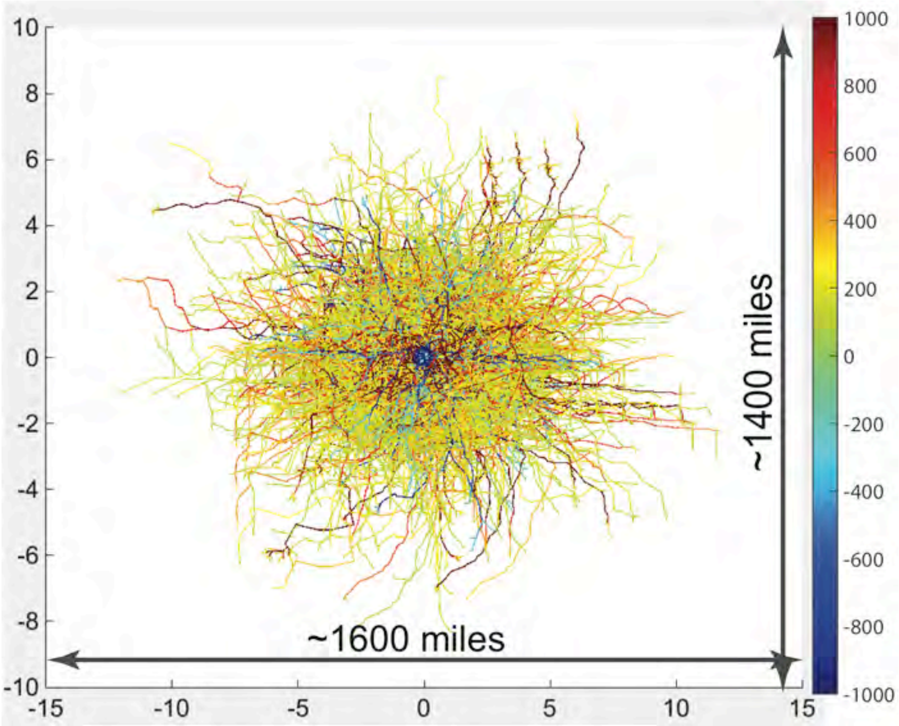


Fig 19: Overlaid traffic changes after a geo-localized WMD events in 1000 uniformly chosen epicenters in the US. While most traffic changes are short-ranged, there are certain epicenters where significant changes can be found far away from the epicenter.

where the index B indicates the case when both the population and the network was removed from the damaged area, index P indicates when only the population was removed but the network was left intact, ΔT_{ij} is the traffic change (due to the damage event) on road segment ij and the summation is over all the road segments in the network. The ratio σ indicates by how many times the effects of damage that include the network is stronger than just due to the removal of population. A larger than unity value for σ indicates a network effect. The larger this ratio, the stronger is the network effect. Figure 20a shows σ for 556 damage centers chosen uniformly from the network, ordering the damages centers on the x-axis according to their σ values increasingly. While we can see that network effects are almost always there, the end tail of σ picks up rather sharply for about 46 damage centers with a σ of 20 or larger. These damage centers of nodes form the fragility points of the network. Figure 20b shows the cumulative distribution of σ, which obeys a power-law decay.

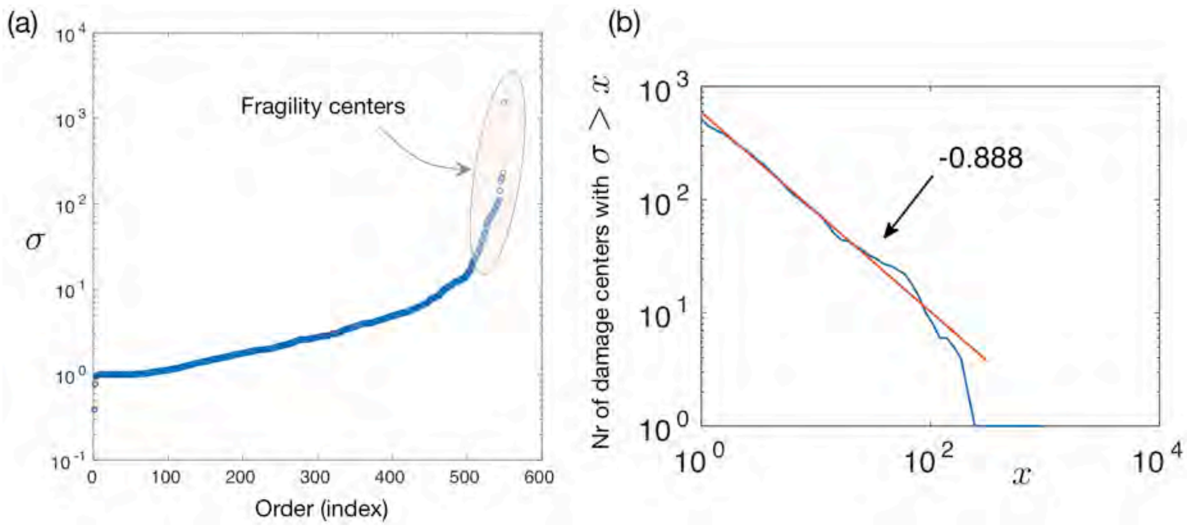


Fig 20: Separating network effects from population effects in damage consequences. (a) Shows that σ sharply increases for a small number of damage centers indicating the existence of network fragility points. (b) Shows the number of damage centers with a network effect ratio σ larger than x vs. x.

The significance of long-range, purely network-effects generated consequences of a geo-localized WMD are shown in Figure 21. In Figure 21a,b we show the case of San Antonio, for removal of both network and population (b) and just population (a). There are no major differences between the effects in the two cases, showing that the changes in this location would be a population effect. However, in the case of La Porte, Figures 21c,d the effects are highly local if only population is removed within the area (c), but when the network is also removed, the effects are significant and spread far out (d).

The patterns of significant flow changes around a damaged area are similar to propagation of cracks in brittle materials, showing tendril-like structures, which then reach far into the rest of the network (Figure 19). As we will show next, one can connect the nature of these effects to Laplacean growth, which also describes crack-propagation among many similar pattern formation phenomena (dielectric breakdown, viscous fingering, etc.) Such phenomena are characterized by power-law scaling, which is also demonstrated for our case, in Figure 22. This figure shows the average value of the maximum traffic change for road segments at a given distance from the epicenter of the damage as function of the distance. The decay is clearly slow, a power-law, showing the existence of long-range effects. We have also developed a method that can identify high fragility points in the network. The algorithm identifies them within structural holes of the network in which there are regions of high population surrounding an area with few access paths within the surrounded area. These results are being written up and submitted for publication.

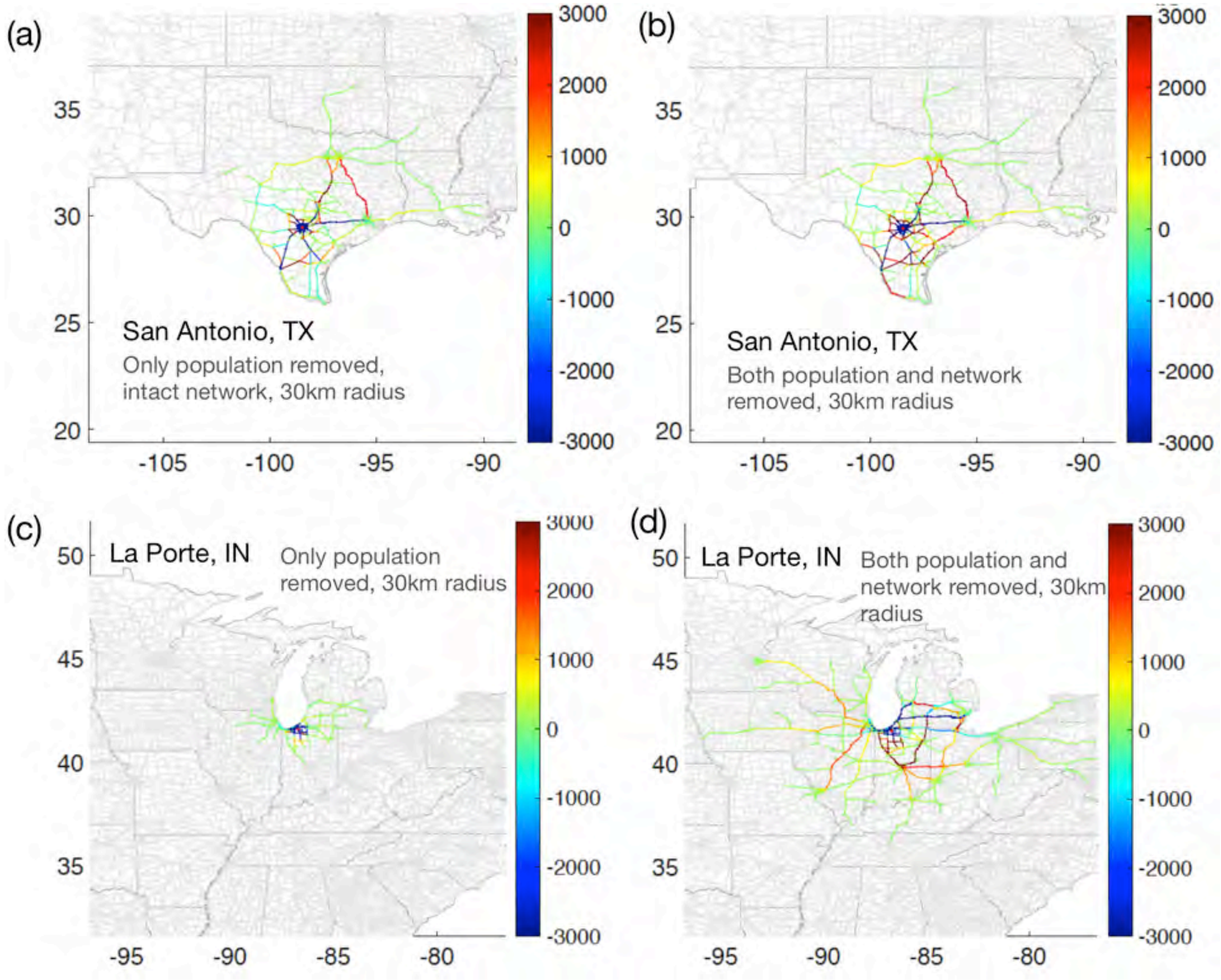

(a)
San Antonio, TX
Only population removed, intact network, 30km radius
(b)
San Antonio, TX
Both population and network removed, 30km radius
(c)
La Porte, IN
Only population removed, 30km radius
(d)
La Porte, IN
Both population and network removed, 30km radius
35
30
25
20
50
45
40
-105
-100
-95
-90
-85
-80
3000
2000
1000
0
-1000
-2000
-3000

Fig 21: Comparison between changes in network flows between cases when only population is removed vs both population and network transportation is removed within the affected area (30km radius). (a-b) case of San Antonio, TX, showing a population effect. (c-d) Case of La Porte IN shows a strong network effect as it is a network fragility node.

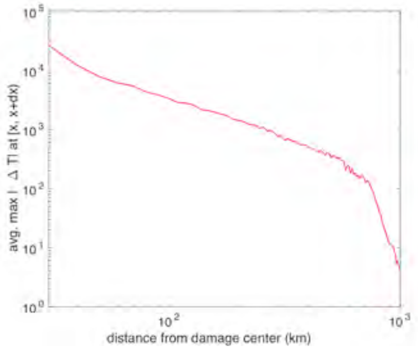


Fig 22: The decay of the maximum traffic change within a ring of about 5km as function of distance from the damage center is power-law indicating the existence of long-range effects. The traffic changes were averaged over 1000 damage centers.

5) Mathematical laws characterizing the non-local propagation of localized damages

We have shown previously via a first principles based modeling approach, the features of the observed flows in this network are determined only by a few fundamental factors, namely the distribution of population, the network structure and a cost-minimization principle. For that reason, we expect that our approach will be applicable to more general flows and network types where these factors are present. We have generalized the radiation law introduced by Simini et al. [1], which in its original form did not couple human dynamics with network structure. Our model, that is the generalized radiation law [1] is based on first principles that lie at the heart of the interactions in the system determining the flows, rather then specific use-cases and fitting parameters. This is very important for modeling the consequences of WMD events on infrastructure transportation networks for the reason that in the wake of a WMD event we can no longer use the same fitting parameters on the modified network that were obtained from a normal operating network behavior. That is why we need first principles based models, which will work on any network, without the need for uncontrolled or poorly understood fitting parameters. While first principles based models will not capture every detail of the system, they do capture the overall behavior and the essential mechanisms behind the overall behavior. Our contribution was to provide a generalized form of the radiation law coupling it with the network structure (using a fundamental principle, namely, cost-minimizing path choice), and demonstrated its applicability for traffic flows using US highway data.

Robustness of the generalized radiation law for transport.

We have also been developing a mathematical understanding of the robustness of the generalized radiation law and the universal character of the scaling behavior observed in the real-world traffic data during last year's work. The scaling behavior shows that the population flux Φ_{ab} from an origin *a* to a destination *b* scales as the inverse second power of the population s_{ab} within the domain that can be reached with transportation costs not larger than the cost of transportation from *a* to *b*: $\Phi_{ab} \sim (s_{ab})^{-2}$. This scaling with population is then responsible for the distribution of the fluxes obeying a scaling law that holds over seven orders of magnitude as reported in our paper from last year analyzing the real-world traffic data. We have shown the robustness of this scaling law by studying a noisy version of the radiation law in which the choice of an agent to travel to a destination site is modified from a Heaviside step function (always accepts the choice if the conditions are met) to a probabilistic version in which the acceptance probability is given by the well-known Metropolis rate $\min(1, e^{-\beta \Delta z})$, where $\Delta z = z - z'$, where *z* is the demand value of the agent and *z'* is the value offered by the destination site. Here β is a parameter that is used to tune the degree of uncertainty in the choice acceptance (it is called inverse temperature in statistical physics). Performing the calculations with this modified model, called the *noisy cost-based radiation law* we find that the probability for a single agent will have to travel from origin *a* to destination *b* to meet his demand is given by:

$$P(1|m_a, n_b, s_{ab}) = \left(1 + \frac{1}{\mu}\right) \frac{m_a n_b}{(m_a + s_{ab})(m_a + n_b + s_{ab})} - \Gamma(1 + \mu) \frac{m_a n_b}{{s_{ab}}^{1+\mu}} \tag{8}$$

where $\mu = \beta / \lambda \gg 1$, assuming a parametric distribution for the demand/offer value z in form of $p(z) = \lambda e^{-\lambda z}$ (other forms lead to similar results). Thus even in the noisy case we see that the leading behavior for this probability (which, when multiplied by the size of the population at the origin site m_a gives the flux Φ_{ab}) is dominated by the inverse square law behavior, the change appearing at the level of small corrections, showing that indeed this is a robust law that can be used in modeling scenarios.

Laplacean models of infrastructure networks.

Previously we have set the goal to explore the hypothesis that the network structure of roadway transportation conforms mathematically to Laplacean growth and the corresponding damage patterns as well. Many systems although different in their composition and detailed behaviors are found to be described by the same mathematics. In particular, the viscous fingering patterns grown in the Hele-Shaw cells or patterns by electro-deposition or mineral deposits or dielectric break-down - all have the same underlying mathematics, namely Laplacean growth. Knowing the mathematical characteristics of infrastructure networks will allow a deeper understanding for their protection and mitigation of the effects of WMD events or other catastrophic events. Diffusion limited aggregation (DLA) is a physical model based on a stochastic process to simulate Laplacean growth phenomena [8], [9]. The mathematics behind all these phenomena is governed by the Laplace equation: $\nabla^2 \varphi = 0$, where φ is a potential field that governs the growth of the patterns with the corresponding boundary condition at the pattern's location (which is changing in time due to growth). We have observed structural similarities between the roadway network and DLA, both showing similar fractal patterns and having same fractal dimension D=1.73. For the roadway network in Fig 23a) we colored the road intersections on the roadways according to the distribution of the travel costs they can be reached from some fixed arbitrary site. Since, as we have shown previously, time-based travel costs reproduce with excellent agreement the flow patterns, we restricted our studies to this cost measure. The two have strikingly similar fractal patterns, however, the autocorrelation function of these patterns are different. The autocorrelation is the correlation between the pattern and itself with a shift by distance r, as function of r. The decay of the autocorrelation function for DLA is a power-law with an exponent of -0.337 as reported by others and reproduced by us as well. On the other hand, the autocorrelation function decay for roadway network is a power law with an exponent of -0.145. However, we could resolve this discrepancy with the observation that the roadway network is a coalescence of multiple DLA patterns. We implemented the multiple-site DLA simulations using CUDA programming language on a Graphic Processing Unit (GPU) and significantly reduced the computation time. The main difference, however between DLA patterns and the roadway network is that the roadway network is a mesh of lines, whereas DLA-s are clusters of points. However, if we transform the roadway network into its line-graph, we find that the corresponding structure can now be compared to DLA type patterns. We have been exploiting this observation and finalizing a mathematical model for the evolution of infrastructure networks. We expect that the same mechanism will describe other large-scale infrastructure networks as well, in which network growth is influenced by population distribution, population growth and cost-minimization for transport.

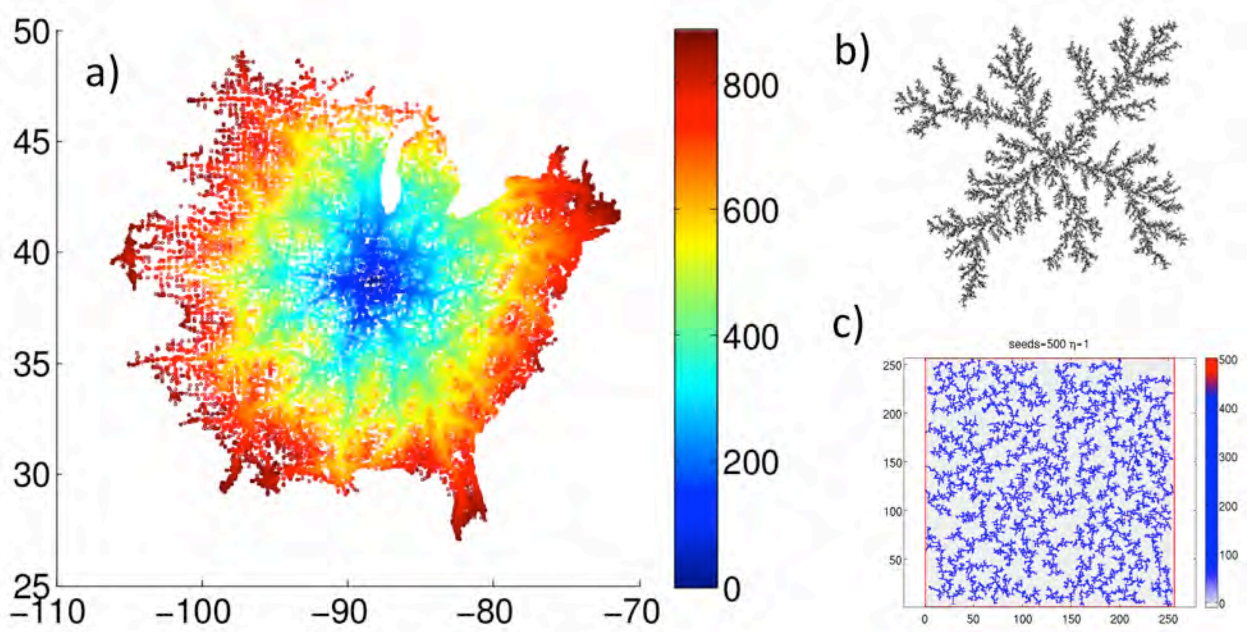


Fig 23: a) Regions reachable within a constant travel time based cost in the US highway transportation network. b) a Diffusion Limited Aggregation (DLA) pattern obtained from the Laplacean growth equation using a single seed. c) Same as in b) from multiple seeds. The generated patterns are structurally similar with same fractal dimension 1.73.

Network vulnerability and shattering.

In parallel, we have continued our research on the vulnerability backbones of infrastructure networks. As defined before by us also as part of this project [6], [7], the vulnerability backbone of a network is the smallest percolating/connected subgraph whose removal minimizes the size of the largest component in the graph remaining after the removal. One can show that finding the vulnerability backbone is an NP-hard problem, similar to computing the toughness measure of the graph [10]. Thus, we must work with heuristic algorithms. In the last period we were able to show that an equivalent formulation can be given to the vulnerability backbone problem using a weighted betweenness approach. The idea is to consider a set of edge-weights $\mathrm{W} = \{w_1, \dots, w_M\}$ and the associated weighted betweenness values on all M edges. Recall that the weighted-betweenness of an edge (node) is the fraction of all-pair lowest-cost paths (where path cost is computed from the sum of weights along its edges) that run through that edge (node). In [7] we provided an algorithm (O(NM)) that computes these betweenness values for all edges and nodes. Next, we consider the set of edges (nodes) whose betweenness value is larger than a given threshold B. Starting from the largest value for the threshold (the largest betweenness value in the network) we gradually lower the threshold and identify the edges (nodes) whose betweenness is larger than the threshold. There will be a critical threshold value B_C at which point the subset of edges (nodes) with betweenness larger than B_C will join into a connected giant cluster, percolating through the network. We then remove this largest cluster and we do a component distribution analysis on the remaining parts of the network. This procedure shatters the network and we are looking to minimize the size of the largest remaining component. The procedure itself is polynomial and thus computationally feasible. The NP-hard part was translated

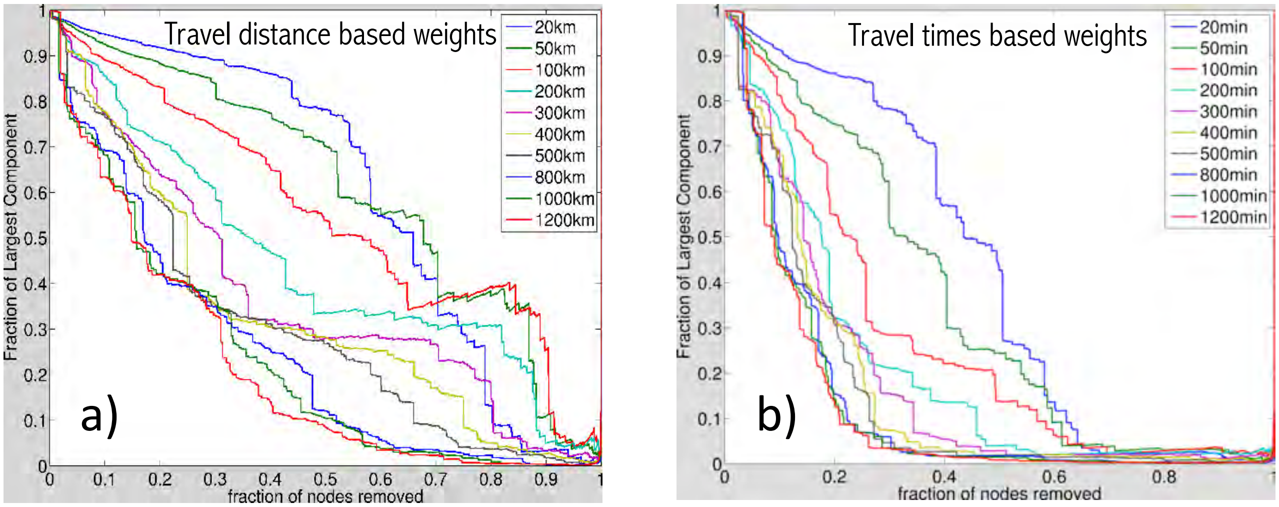


Fig 24: Network vulnerability backbone detection and shattering performance. We used a weighted-betweenness approach to detect and remove the vulnerability backbone using a) distance-based weights and b) travel times based weights in the US highway transportation network. The travel-time based distances provide a much better performance to network shattering.

to the assignment of the weights $W = \{w_1, \dots w_M\}$. The problem is then finding a weight assignment such that the size of the largest remaining component after the removal is as small as possible (efficiently disconnecting the whole network). We tested several weight assignments, but what we have found is that the weights obtained by assigning travel times on each road segment has led to an optimal performance for the shattering algorithm. In Fig 24 we show a comparison between two cases. In a) we used travel distances as weights, whereas in b) we used travel times as weights. On the y-axis we show the fraction of the largest connected component after removing a given fraction of the nodes with the largest weighted edge-betweenness for that weight assignment. The different colored curves correspond to different ranges within which the node pairs were taken and their contribution to edge-betweenness calculated [6], [7]. The faster and the more significant the drop of the curves, the more efficient is the shattering.

As we can see, the time-cost based weights provide a significantly better performance than the distance-based weights. In particular, by removing about 20% of nodes we could shatter the network such that after the shattering the largest connected component contained less than 10% of the nodes. In the distance-based weights case removal of 20% still leaves a 40%-large connected component. Why are travel-time based weights so efficient in solving this NP-hard global optimization problem? The reason lies with the fact that human travel tends to optimize travel times and the speed limits are correlated with usage demand. Similar to ant-based optimization algorithms, human mobility extracts the structural information necessary for efficient travel and adapts to it. In this sense, we are extracting the subgraph for optimal shattering/separation from monitoring usage efficiency of the network paths. Figure 25 shows the removed subgraphs (red) at various 15% and at 30% threshold of the largest weighted betweenness values. The method identifies the high traffic/transport backbone of the network.

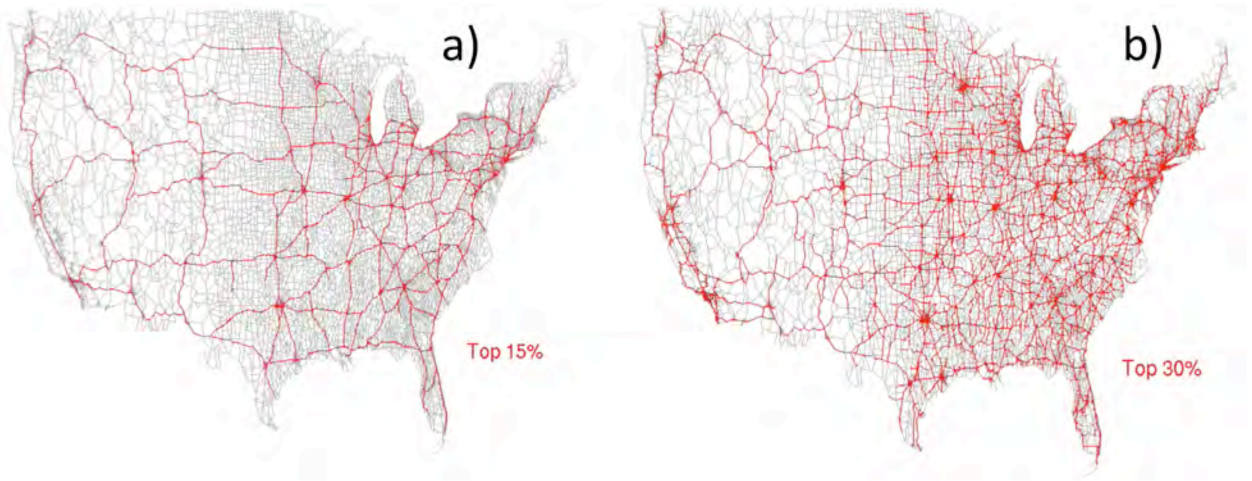


Fig 25: Top 15% a) and top 30% b) highest weighted betweenness subgraphs of the US highway transportation network. The removal of these subgraphs generates the largest shattering effect on the network corresponding to Figure 24b).

6) An E-WMD threat through the International Food Transportation Network (IFTN)

The weighted betweenness centrality based method to compute the vulnerability backbone of a network introduced by us and used for the US Highway Transportation Network as discussed above can be applied to transportation networks in general. To test this, and thus further validate this method, we applied it in a potential E-WMD type scenario, namely spreading bio-pathogens and/or toxins through the international food transportation network (IFTN). In collaboration with food science experts from UK and Hungary we have obtained and analyzed data from the UN's statistics division, about the amounts of food (dollar value) imported and exported globally, amongst 207 countries within a period of 10 years, 1998-2008. Our analysis of this data [11] has shown an amazing complexity underlying this transport network (see Fig 26), a dense network of trades making it extremely difficult to trace the origins of any food-borne (accidental or intentional terrorism related) diseases. By applying our weighted betweenness approach [7] we were able to not only identify the most vulnerable backbone of this network, but also derived a contaminant flux model (Fig 27), the results of which are in very good agreement with known major cases of food-borne disease outbreaks worldwide (Fig 28). Accordingly, the largest mass effect would be achieved if pathogens were introduced in food shipments from the Netherlands going through Germany. Such effects are not only in terms of the number of people infected and

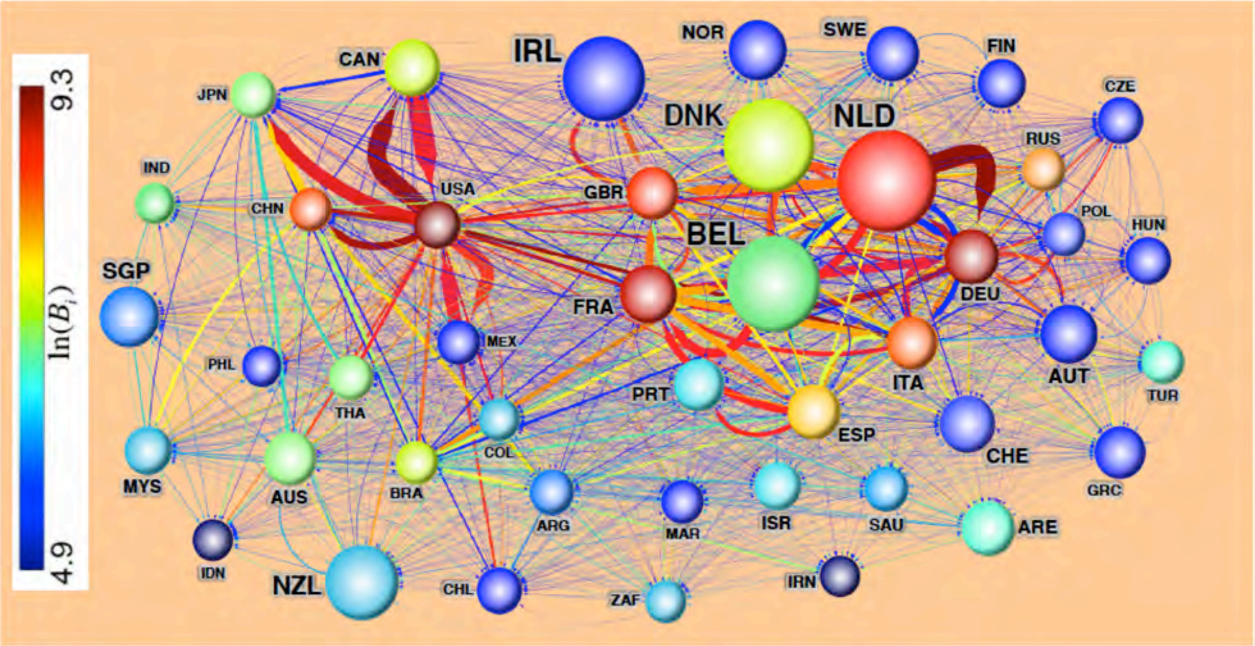


Fig 26: The *backbone* of the International Agro-Food Trade Network (the full network is too complex for graphical representation) of 44 countries identified with our weighted betweenness centrality method. There are 7 countries here that are responsible (each) for trading with at least 77% of all other countries and form the *vulnerability* backbone, from where contaminants would spread most efficiently into the rest of the network. Nodes and edges are both colored by the logarithm of their betweenness values (see color-bar); the thickness of the directed edges is proportional to the logarithm of the trade value in that direction.

loss of life, but also in the major political and economical repercussions that follow. A case in point is the 2011E.coli O104 outbreak in Germany http://blog-admin.wired.com/wiredscience/2011/07/e-coli-3-years/, with almost 4000 illnesses and 44 deaths. A detailed study taking several months identified the source of illness in form of seeds shipped from a port in Egypt which then "took a tortuous path" through Europe, first through Belgium, then The Netherlands (Rotterdam) then Germany.

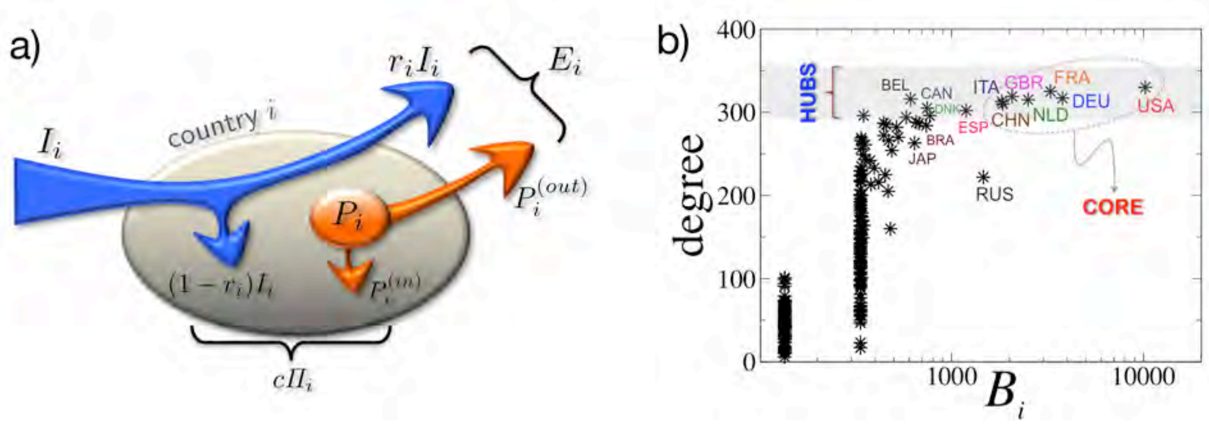


Fig 27 : a) The Contaminant Food Flux Model: Country i with population of P_i has a total yearly agro-food import I_i, out of which r_iI_i is exported, and $(1 - r_i)I_i$ is consumed locally. A specific food ingredient to be tracked is produced in this country in the value of P_i from which $P_i^{(out)}$ will be included into its total export E_i, while $P_i^{(in)}$ is consumed locally. The parameter c represents the average value (in US$) of the food consumed by a person in a year. b) A scatter-plot of degree vs. betweenness for every country.

Note that our study does not predict an increase in the number of food poisoning cases but that, when it happens, there will be inevitable delays in identifying the sources due to the increasingly interwoven nature of the IFTN. That is, even if food contamination was less frequent (for example due to better local control of production), its dispersion or spread is becoming more efficient. In particular, our study identifies critical spots in the network that may seriously hamper future bio-tracing efforts. Although the analysis presented here is based on coarse data representing aggregated food fluxes, it can also aid with bio-tracing, in a "Bayesian approach" sense by providing a list of most

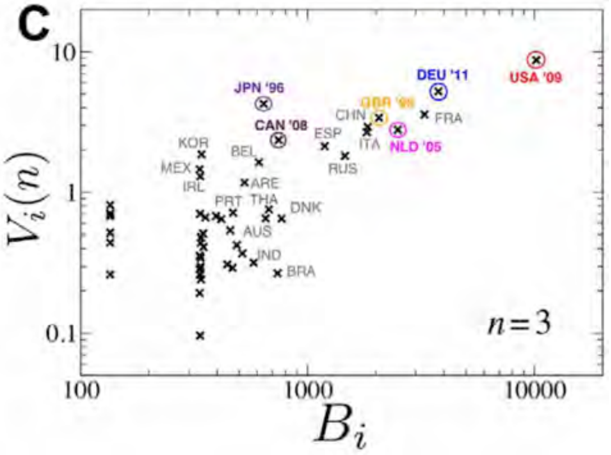


Fig 28: "Vulnerability vs. betweenness" scatter plot for the 44 countries with the largest trade activity. Countries with significant food poisoning cases in the last 15 years are indicated by encircled symbols. In particular: the 2011 Listeria outbreak in the USA, from produce, causing 29 deaths; the 2011 E. coli outbreak in Germany, from red beet sprout, with 46 deaths and 4000 diagnosed cases; the Salmonella outbreak in 2005 in The Netherlands with 165 diagnosed cases; the 1996 E. coli outbreak in the UK with 512 confirmed cases, 17 deaths; the 2008 Listeria outbreak in Canada with 57 diagnosed cases and 27 deaths; the 1996 E. coli outbreak in Japan, from radish sprout, with 2 infant deaths and more than 5000 hospitalized.

probable sources and pathways to be used as starting points. Our findings were published in the journal PLoS ONE [11] with the paper receiving over 3200 views in the first 6 weeks, and several news outlets have published feature articles on it.

7) Expanding the predictability range for the Maximum Entropy (MaxEnt) Principle

When we are trying to understand a system or predict its behavior, we usually do that based on limited information. In everyday situations we make our decisions and choices using some sort of "inner" probability distribution shaped by past experience. However, human decision-making tends to be biased; the wiki page [12] enlists over 150 such biases. Naturally, this raises the question whether there is a way to make in-silico, unbiased quantitative inferences based on limited information.

The founder of a principled and quantitative approach to unbiased inference making is physicist Edwin T. Jaynes. Standing on the shoulders of Laplace, Bayes and Shannon, Jaynes started a revolution in statistical inference with his two seminal papers from 1957 [13] [14]. He based his inference method on the notion of statistical ensembles, much in the same way that statistical mechanics uses the ensembles of microstates to describe macroscopic properties of matter.

The MaxEnt is based on a simple premise: Given the available information about a system, make predictions using only that information and without introducing additional biases. Tautologically speaking, once the available information has been incorporated, there is no information left for us to exploit. It is equivalent to predicting the throw of a fair coin, or the upside of a thrown fair dice. Clearly the uniform choice will be the best choice in this case, but to express this in a general and quantitative way, following Jaynes, one uses the notion of information theoretic entropy introduced by Claude Shannon. Let Ω denote the space of all possible states of our system of interest. A particular state (which can also be conceived as an "event") $\vec{\mu} \in \Omega$ of the system (or microstate in physics parlance) specifies all the variables necessary to uniquely determine the state. For our purposes it is sufficient to think of Ω as a discrete collection of states, but in general can be a continuous space as well. Our predictions about the system will be in the form of a probability distribution $P(\vec{\mu})$ over Ω. One can think of the available information as *constraints* on the set of microstates that the system can be in: without any information we cannot say anything about the state of the system and thus any state is fair game with the same probability, i.e., $P(\vec{\mu}) = |\Omega|^{-1}$and thus the uncertainty is maximum. As information is added, it changes the distribution to something else; if we happen to know the precise state of the system $\vec{\mu}_0$, then $P(\vec{\mu}) = 1$ for $\vec{\mu} = \vec{\mu}_0$ and $P(\vec{\mu}) = 0$ for $\vec{\mu} \neq \vec{\mu}_0$, and in this case there is no uncertainty. Shannon defines information as the amount of *surprise* we get when some event actually *happens*, whose probability of happening is $P(\vec{\mu})$. Without going into details, the information content in the *event* $\vec{\mu}$ happening is defined as $I(\vec{\mu}) = -\log P(\vec{\mu})$. If $\vec{\mu}$ is the event of snow falling in San Diego then when that actually happens is very surprising (large $I(\vec{\mu})$) and even news agencies will talk about it, $P(\vec{\mu})$ is very small (last happened in 1967). The same event happening at the North Pole, however, carries no surprise ($P(\vec{\mu})$ close to unity, small $I(\vec{\mu})$). As a functional of the distribution $P(\vec{\mu})$, Shannon's entropy

$$S[P] = \langle I(\vec{\mu})\rangle = -\sum_{\vec{\mu}\in\Omega} P(\vec{\mu}) \log P(\vec{\mu}) \tag{9}$$

expresses the average uncertainty induced by the *distribution* $P(\vec{\mu})$. The available information is then inserted in form of *ensemble averages* of observables $m_i(\vec{\mu})$, $i = 1, \cdots, K$ (here K is the total number of constraints):

$$\overline{m}_i = \sum_{\vec{\mu}\in\Omega} m_i(\vec{\mu}) P(\vec{\mu})\,, \quad i = 1, \cdots, K\,. \tag{10}$$

The $\overline{m}_i$ represent values that we know about the system (e.g., we measured them, someone gave them to us, or we simply assumed them). Thus, we are looking to find that distribution $P(\vec{\mu})$ which obeys the constraints (10), it is normalized $\sum_{\vec{\mu}\in\Omega} P(\vec{\mu}) = 1$ and *beyond that* it is the least biased, or maximally uncertain, i.e., for which entropy (9) is maximal. As Jaynes formulates it, this is the distribution that is the least committal towards the unavailable information. Using the standard method of Lagrange multipliers we then maximize (9) subject to (10) and normalization, and in the end obtain the parametric family of Gibbs distributions:

$$P(\vec{\mu}; \vec{\beta}) = \frac{1}{Z(\vec{\beta})} e^{-\vec{\beta}\cdot\vec{m}(\vec{\mu})} \tag{11}$$

where

$$Z(\vec{\beta}) = \sum_{\vec{\mu}\in\Omega} e^{-\vec{\beta}\cdot\vec{m}(\vec{\mu})} \tag{12}$$

is the normalization factor, also called *partition function*. The $\vec{\beta} = (\beta_1, \cdots, \beta_K)$ Lagrange multipliers are then determined from solving (10) with the form of (11) replacing the $P(\vec{\mu})$ to obtain unique values $\vec{\beta} = \vec{\beta}(\overline{m}_1, \cdots, \overline{m}_K)$. This step is also called fitting because these are usually complicated equations that can only be solved numerically.

<u>An example: Throwing 3-sided dice</u>

As a simple example, consider a 3-sided dice. (On a standard dice repeat every number 1,2,3 twice on its opposite sides) Thus $\Omega = \{1,2,3\}$. Assume that after throwing the dice very many times Alice reports to Bob the average value of the upside face as $\overline{m} = 2.5$. Based on this information only, Bob wants to compute the MaxEnt probabilities $P(1)$, $P(2)$, $P(3)$ of the topside face. In this case equation (10) is $P(1) + 2P(2) + 3P(3) = \overline{m}$ and with normalization $P(1) + P(2) + P(3) = 1$ we have one more unknowns than equations. The MaxEnt distribution will simply be $P(m;\beta) = e^{-m\beta}$, $m = 1,2,3$ and $Z(\beta) = e^{-\beta} + e^{-2\beta} + e^{-3\beta}$. Equation (10) becomes a quadratic equation for $e^{-\beta}$ with the only valid solution of:

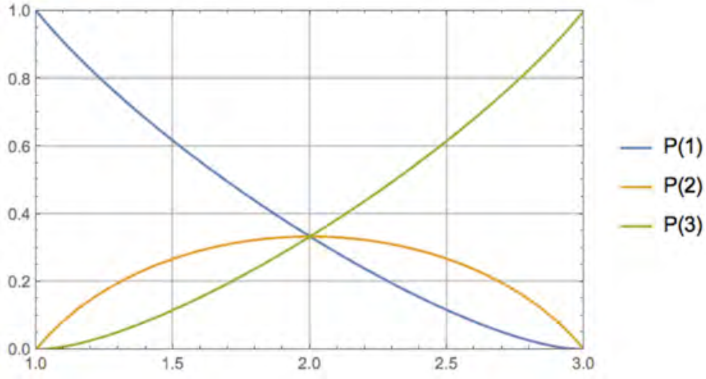


Fig 29: MaxEnt probabilities for a simple example.

$$e^{-\beta} = \frac{\overline{m} - 2 + \sqrt{\Delta}}{2(3 - \overline{m})}\,, \qquad \Delta = -3\overline{m}^2 + 12\overline{m} - 8 \tag{13}$$

leading to $P(1) = \left(10 - 3\overline{m} - \sqrt{\Delta}\right)/6$, $P(2) = \left(\sqrt{\Delta} - 1\right)/3$ and $P(3) = \left(3\overline{m} - 2 - \sqrt{\Delta}\right)/6$. In particular, for $\overline{m} = 2.5$, we get $P(1) \cong 0.116$, $P(2) \cong 0.267$ and $P(3) \cong 0.617$, indicating the most likely biases in the dice. The plots of these three probabilities as function of $\overline{m}$ are shown in Fig 29.

What if Alice reports $\overline{m} = 2$? In this case $P(1) = P(2) = P(3) = 1/3$, which are the *same* values as for an unbiased dice. Does this mean that the dice is fair? If the only information we have is the average topside face value $\overline{m}$, we cannot say whether the dice is biased, only that the available information does not contradict the assumption that it is a fair dice. To see this, assume that Alice reports in addition, also the average of the square of the upside face value, i.e., $\overline{m^2}$. In this case we have full information, namely, as many unknowns as independent equations and we find that $P(1) = \left(6 - 5\overline{m} + \overline{m^2}\right)/6$, $P(2) = 4\overline{m} - \overline{m^2} - 3$, $P(3) = \left(2 - 3\overline{m} + \overline{m^2}\right)/2$. Thus, if Alice reports that $\overline{m} = 2$ (same as it would be for a fair dice) but $\overline{m^2} = 5$ then $P(1) = P(2) = 1/2$ and $P(3) = 0$, and we actually have a heavily biased dice. For a detailed account on the MaxEnt method, its dynamical version MaxCal and for their many applications see the review in Ref [15] by Pressé et al.

<u>The degeneracy problem</u>

We first illustrate the degeneracy problem on a fictitious social network example: Let us assume that intelligence agencies have discovered that a new extremist but nefarious and intolerant ideology is secretly taking hold across the globe. From indirect evidence and intercepted, but heavily encrypted communications it was found that there are $N = 9$ cells involved in terrorist activities all over the world under the aegis of this new ideology. Based on their data, the defense analysts expect that $\bar{M} = 17$ pairs of cells interact regularly and that $\bar{T} = 19$ different triplets of cells have collaborated at various times to realize their attacks. However, it is not known specifically which cells are interacting with one another and which cells participated in which activity at any time. The analysts want to know the likelihood that the 9 cells form a single connected network and they also want to find the most likely network that the available information $(N, \bar{M}, \bar{T})$ is consistent with.

This problem seems ideal for the MaxEnt method. In this case the set of microstates Ω is formed by all the simple, labeled graphs $G \in \Omega$ (here G replaces $\vec{\mu}$ in the above) on $N = 9$ nodes. There are $|\Omega| = 2^{N(N-1)/2} = 68719476736$ such graphs in total. The two constraints in expectation are the number of edges $\bar{M} = 17$ and the number of triangles $\bar{T} = 19$. Since the number of nodes is small, in this case we can exactly enumerate (*in-silico*) all the simple, labeled graphs (graphs from here on) with a given number of edges and triangles and thus compute exactly all the probabilities involved in Jaynes's method. If β is the multiplier corresponding to the number of edges and γ for the number of triangles, after solving (10) for β, γ one finds $\beta = 1.3017$ and $\gamma = -0.6325$ and the rather peculiar distribution shown in Fig 30.

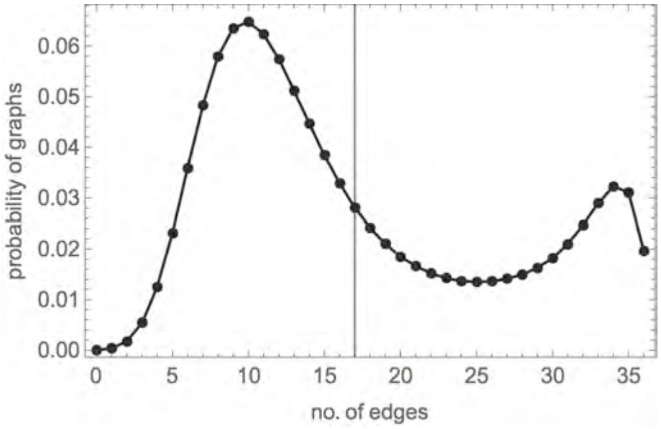


Fig 30: Distribution of graphs with given number of edges in the MaxEnt model for the terrorist network of cells.

It seems as if the MaxEnt method is confused and cannot make up its mind: the microstates (graphs) with the highest probabilities (two peaks in the figure) are either very sparse (few edges) or very dense (almost all connections are there). The average number of edges is certainly respected from the combination of the two types in the sum (10) but the individual realizations/microstates are rarely near the expected value. Moreover, in one case (sparse) the network is disconnected into pieces, whereas in the other (dense) forms a single component, and both are highly probable. But the situation is even worse: since in this case we can enumerate everything, we find that there are 64260 (labeled) graphs with exactly 17 edges and 19 triangles but *none of them* is fully connected. Yet, the MaxEnt method tells us that there are a significant number of connected graphs! More precisely, the probability that a graph sampled by the MaxEnt method will be fully connected is 0.6. None of the connected graphs (of the 60%), however, has 17 edges and 19 triangles; they don't have the observed property at all. This answer is completely useless for our analysts. This effect is called *degeneracy*, and has been known for nearly three decades, since the works of David Strauss [16] in social networks. Other examples of degeneracy are shown in our paper [17], one involving the well-known karate-club dataset by constraining the number of edges and two-stars, another involving a network of jazz musicians, where we constrain three quantities in expectation: edges, two-stars and triangles and a third example involving groups of permutations (related, e.g., to predicting outcomes of horse races). Note that the constraining values $(\bar{M}, \bar{T})$ could have come from an actual realization/state of the system from Ω. A degenerate distribution in this case may even predict a near zero probability for the realization from which the data was obtained! Historically, in degenerate cases people simply have thrown away the existing data and tried choosing other types of measures and data for it. With some luck, degeneracy would be avoided. This is certainly unsatisfactory, as we might not be able to collect new data on other types of measures (as in the case of the nefarious network of cells above), or the measures used are of particular interest in the field (like triangles in sociology). We thus have to have a method that maximally exploits the data that is available to us, instead of avoiding them.

What causes degeneracy in the MaxEnt method?

To answer that, let us first make the observation from (11) and (12) that all microstates (graphs in our case) that share the same values for the measures $\vec{m}$ will all have the *same* MaxEnt probability. So instead of summing over the space of microstates (which is impossible because it is huge and usually we know very little about it), we can gather all the terms (microstates) in the sum with the same $\vec{m}$ into one group and then sum over the range of values for $\vec{m}$. In particular, the partition function (12) becomes:

$$Z(\vec{\beta}) = \sum_{\vec{m}} \mathcal{N}(\vec{m}) e^{-\vec{\beta}\cdot\vec{m}} \tag{14}$$

where $\mathcal{N}(\vec{m})$ is the number of microstates with given values for the measures $\vec{m}$. In physics this function is called the density of states (d.o.s) function. Thus, the averages (10) can be expressed as

$$\bar{m}_i = \sum_{m_i=-\infty}^{\infty} m_i \, p(\vec{m}; \vec{\beta}) \qquad i = 1, \cdots, K\,. \tag{15}$$

where

$$p(\vec{m}; \vec{\beta}) = \frac{\mathcal{N}(\vec{m})}{Z(\vec{\beta})} \, e^{-\vec{\beta}\cdot\vec{m}} \quad . \tag{16}$$

Note that in all the above the $\vec{m}$ appear as free variables. Eq. (16) expresses the probability that the MaxEnt distribution generates (or samples) a microstate with given values for measures $\vec{m}$. Thus, in degenerate cases $p(\vec{m};\vec{\beta})$ shows two or more peaks as function of $\vec{m}$ (see our paper for a more precise description). Since the exponential in (16) is a monotonic function of $\vec{m}$ it then follows that the culprit behind degeneracy is the d.o.s function $\mathcal{N}(\vec{m})$. In the paper we prove the following theorem:

Theorem: A MaxEnt model is non-degenerate if and only if the density of states function $\mathcal{N}(\vec{m})$ is log-concave.

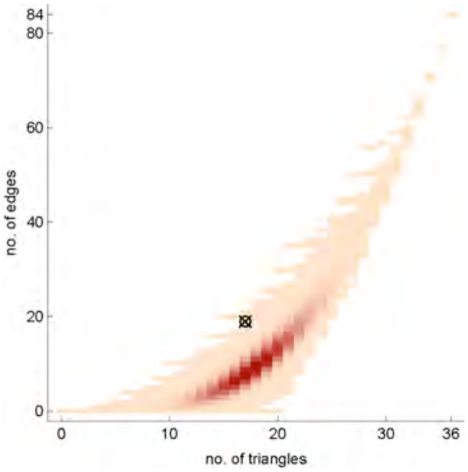


Fig 31: The d.o.s. function on 9 nodes (color intensity) with given number of edges and triangles. The cross hair is at 17 edges and 19 triangles.

A function is log-concave if the log of the function, that is, $\log \mathcal{N}(\vec{m})$ is concave. The fact that one can pinpoint the cause of degeneracy is the good news. The bad news is that it is related to properties of the function $\mathcal{N}(\vec{m})$. It is a *counting* function, giving the number of objects in some space that all share the same property, for example in our case the number of simple graphs on N nodes with a given number of edges and a given number of triangles (precisely that many, not on average!). This is a purely mathematical/combinatorial object and has nothing to do with MaxEnt, probabilities or inference, however, its properties determine whether the MaxEnt method will show degeneracy. The bad news is that counting problems are among the hardest problems in discrete mathematics. Even those problems whose decision version is easy (they are in P), their counting version can be very hard (from #P-complete). For example, deciding that there is perfect matching in a bipartite graph is easy (has a polynomial-time algorithm) but counting their number is very hard (no such algorithm).

For a function $\mathcal{N}(\vec{m})$ to be log-concave, two conditions need to be fulfilled. On one hand the domain $\mathcal{D}$ of $\mathcal{N}(\vec{m})$ has to be *geometrically convex* and on the other, it has to fulfill the Prékopa-Leindler inequality. In many cases the first condition is broken, such as in our example of the nefarious network with $\mathcal{N}(\vec{m}) = \mathcal{N}(\vec{m}_{|}, \vec{m}_{\Delta})$ illustrated in Fig 31. Since the domain is curved, waxing-crescent shaped, the domain is non-convex (a straight segment with end-points close to the tips will have most of its points outside the domain). Fig

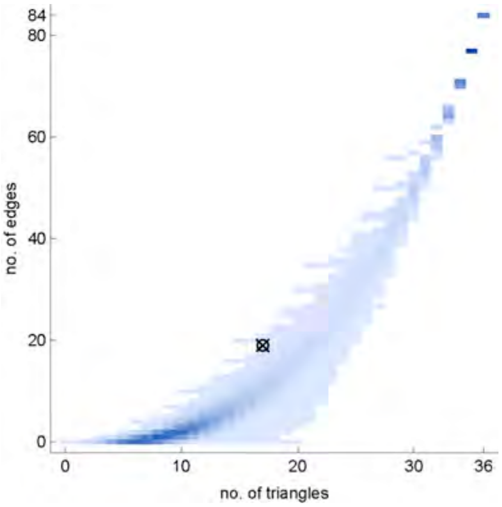


Fig 32: The MaxEnt model samples the graphs according to a bimodal distribution (peaks are at darker blue regions).

32 then shows in this 2D map the bimodal distribution of graphs $p(\vec{m};\vec{\beta})$ as sampled by the MaxEnt method. The non-convexity of $\mathcal{D}$ comes from the fact that the variables, i.e., the components of $\vec{m}$ are not independent, more precisely, they are *non-linearly dependent* on one another. In classical statistical mechanics the constraining variables (in expectation) are independent quantities characterizing different types of interactions with the environment such as pressure (mechanical interactions), chemical potential (chemical or particle exchange), temperature (thermal), magnetic susceptibility (magnetic), etc. In this case the domain of the d.o.s. function is always convex (a hyper-rectangle) and degeneracy does not occur. Note that if the variables are linearly dependent, the domain is still convex and degeneracy does not occur in this case either (assuming the concavity condition for the log of the d.o.s. is valid).

Can we eliminate degeneracy?

Assume that we don't have the possibility to switch to a different set of variables, but instead we have to work with the available measures and the data for them. Still using the same data, and without losing information we would like to make good predictions based on the MaxEnt approach. Let us call our original (degenerate) model (or ensemble) the MaxEnt($\vec{m}$) model (ensemble). The idea is to introduce a *one-to-one mapping* (thus no information is lost) from the original variables (and the data for them) to new ones $\vec{m} \leftrightarrow \vec{\xi} = \vec{F}(\vec{m})$ such that the corresponding d.o.s. function (number of microstates with given $\vec{\xi}$), $\mathcal{N}'(\vec{\xi})$ is now log-concave and thus the new model (ensemble) MaxEnt$(\vec{\xi})$ is non-degenerate. Since $\vec{F}$ is bijective, the number of states with $\vec{\xi}$ is the same as the number of states with $\vec{m} = \vec{F}^{-1}(\vec{\xi})$, i.e., $\mathcal{N}'(\vec{\xi}) = \mathcal{N}\left(\vec{F}^{-1}(\vec{\xi})\right)$. The observation behind our idea is that while a function could be concave (or convex), its composition with another function, in the new variables can have altered convexity properties. Note that such mappings do not require the collection of new data! One simply applies the transformations on the data we have.

While we do not have a general method for finding such transformations, in concrete situations they can be done relatively easily. A simple example of degeneracy in our paper [17] is the case when the constraining variables are the number of edges $m_{|}$, and the number of wedges (2-stars) $m_{\vee}$. Since on average the latter is proportional to the square of the former, we may choose $\xi_{|} = {m_{|}}^2$ and $\xi_{\vee} = m_{\vee}$ (the choice is not unique, see our paper). In our terrorist network example, we may choose $\xi_{|} = {m_{|}}^3$ and $\xi_{\Delta} = m_{\Delta}$. In all these cases the new model MaxEnt$(\vec{\xi})$ will be non-degenerate and based on the same input information with $\bar{\xi}_i =$

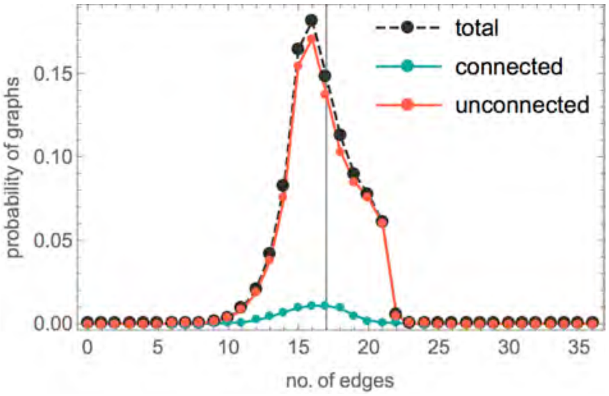


Fig 33: In the corrected model, MaxEnt$(\vec{\xi})$, typical samples are drawn from near the original data values.

$F_i(\overline{m}_1, \cdots, \overline{m}_K)$ (chosen that way). Since MaxEnt$(\vec{\xi})$ is a new model, we may ask how far are the averages of m_i within the new MaxEnt$(\vec{\xi})$ ensemble from the original data values $\overline{m}_i$. Since, for example, instead of requiring that the number of edges on average is 10, we now require that the square of the number of edges is 100 on average—we do not expect this to cause differences that are too big. Indeed, we can show that due to the fact that the new MaxEnt$(\vec{\xi})$ ensemble is concentrated around $\bar{\xi}$ — , the differences are actually rather small (see the last paragraph on pg 3 of our published paper [17]), and the new model still respects (10) with good approximation. Fig 33 shows the corrected MaxEnt distribution by edge-number (black) for the case of the nefarious network. Indeed, the samples are now coming with high probability from around the observation value of 17. The red in Fig 33 shows the probability that a MaxEnt sampled graph with a given edge number will be disconnected and the green that it will be connected; now the MaxEnt predicts that 94% of sampled graphs will be disconnected, much better than previously. Fig 34 shows the MaxEnt distribution in 2D, in the corrected MaxEnt$(\vec{\xi})$ model; it is concentrated around the observation values. Fig 34 is in terms of the original variables; in terms of the new ones it would have a convex shape, see our paper [17].

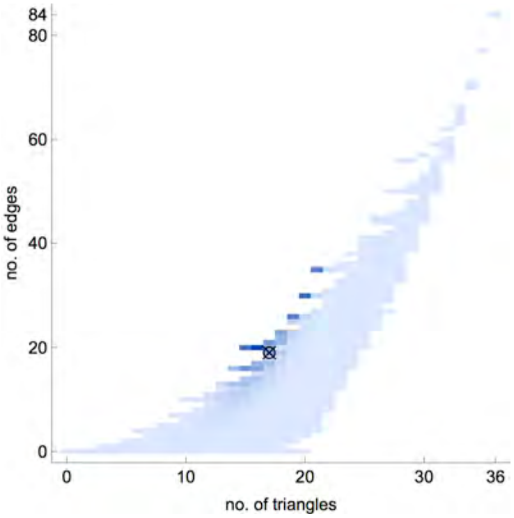


Fig 34: The corrected MaxEnt$(\vec{\xi})$ distribution as function of the complete set of variables.

Discussion

In our work we have shown that the root cause of degeneracy in MaxEnt models is the existence of non-linear constraining relationships between the variables. One can eliminate degeneracy by using a one-to-one nonlinear mapping to a set of new variables (using the same data), and defining a MaxEnt model based on these new variables. The new MaxEnt model will be generating samples with high probability from the neighborhood of microstates for which the given constraints are typical, and the mean values of the original variables in this new model will be close to the original input data. A few comments on MaxEnt models are in order, notwithstanding degeneracy issues.

A. The input data may be applied within a frequentist or a Bayesian framework. The data may come from several repeat observations and the averages represent actual averages of those observations (frequentist). Or, based on our best knowledge we may believe that we can interpret the single dataset available to us as representing typical values and thus we choose to treat them as expected values (Bayesian). In the latter case we will actually generate a MaxEnt ensemble within which the input values are indeed typical. However, if the data came from an atypical or special state of the system, then our predictions will be off when compared against new empirical data for those predictions. When that happens we know that our view of the system is off and thus we can use the MaxEnt as a tool to help update our understanding of a complex system.

B. Given a set of constraining measures $\vec{m}$ we sometimes want to predict the value (in expectation) of another measure (such as the number of triangles knowing the number of edges and 2-stars), or even its distribution. This can only be done to the extent that the constraining measures determine the variable of interest. However, if the variable of interest is independent from the constraining ones, the MaxEnt model will be useless for predicting that variable, because the information we have says nothing about what we are interested in. Jaynes makes the observation that this aspect is actually a strength of the MaxEnt method: if our predictions are way off for a quantity of interest when compared against empirical data, it means that we have not included the relevant variables into our model for that quantity. Unfortunately, this does not say what are the relevant variables, just that the ones used are irrelevant.

C. The output of the MaxEnt model is the distribution (11) and we have to sample microstates from Ω according to this distribution. This, in itself, is a thorny issue and one typically does it using a Markov Chain Monte Carlo algorithm. As a matter of fact, this step is a computational bottleneck that limits in general MaxEnt applications. However, this difficulty can be somewhat circumvented by using proper sampling algorithms.

D. One of the original criticisms of Jaynes's method was its subjective (Bayesian) nature. Namely, physical reality does not depend on our state of knowledge and thus it brought into question the method that maximizing the information entropy (which expresses the uncertainty in *our* knowledge about a system) can be used to generate predictions about physical reality. However, since then several authors, starting with Shore and Johnson have demonstrated that any algorithm that draws inferences about probability distributions in a self-consistent and unbiased way will be maximizing the entropy function (9) — see e.g., [15].

Open problems

MaxEnt based prediction methods are principled and used extensively in applications in practically all fields of natural sciences. However, there are still a number of issues with them (the degeneracy problem we tackled being one of them) that warrant further investigations:

1. Is there a general method/algorithm that computes/generates the mapping $\vec{F}$ when $\mathcal{D}$ is simply connected such that in the new variables the counting function is log-concave?

In 2D, we know from the Riemann mapping theorem that there is a conformal mapping that takes any simply connected domain into the unit disk in the complex plane, which could be applied in this case. What about higher dimensions? Can we define $\vec{F}$ as a computable flow that maps a non-convex $\mathcal{D}$ into a convex $\mathcal{D}'$?

2. Speeding up the sampling (MCMC) process of the microstates according to the MaxEnt distribution (11).

3. Another computational difficulty is solving for the Lagrangean multipliers in (3 (or (15)), because in general we do not know explicitly the counting function. This is usually done approximately, with various stochastic root-finding methods that have various issues, and thus better methods would be desirable.

8) Degree-based modeling

One of the fundamental approaches in network modeling is constructing synthetic networks based on partial data or with statistical characteristics similar to data. Frequently, this data is given on the level of degree sequences or degree distributions. However, given a sequence of degrees, there can be many graphs having that sequence for the degrees of its nodes. For example, Erdos-renyi random graphs and random geometric graphs both have Poisson degree distributions, yet they have very different other properties (clustering, path lengths, etc.). The number of graphs with a given degree sequence is typically a very large (exponentially large) number already for modest size graphs, one cannot possibly construct all of them. Thus in

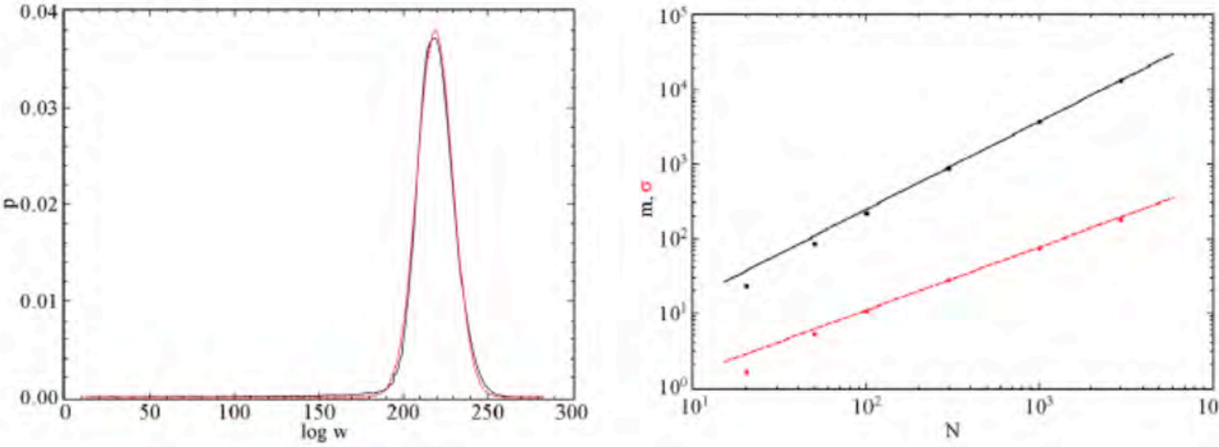


Fig 35: Graph sampling by our direct-construction method for undirected graphs. a) Probability distribution of the logarithm of weights for an ensemble of power-law sequences with $N = 100$ and decay exponent of $\gamma = 3$. The ensemble contained 2×10^4 graphical sequences, and for each sequence 10^6 graph realization samples were produced. Thus, the total number of samples produced was 2×10^{10}. The simulation data is given by the solid black line and a Gaussian fit of the data (corresponding to lognormality) is shown by the dashed red line that nearly obscures the black line. b) Performing the same fits as in a) but for various systems sizes N (same $\gamma = 3$) here we plot the growth of the mean (m) and the width (σ) of the Gaussian as function of N. These show that the spread grows slower than the mean which means that the distribution in the large system limit approaches a Dirac-delta function and thus sampling becomes uniform, asymptotically. A similar behavior has been found also for directed graphs, see [21].

network modeling problems one needs to study network observables over graph ensembles and formulate our answers in statistical terms. In order to avoid biases due to graph sampling in our conclusions, one has to generate uniform random samples of graphs with the given degree sequence or distribution, so that we can assess to what extent a certain observable describing a process or phenomenon (e.g. rate of disease spread) is determined by the given information, in this case the degree sequence. Uniform sampling from graphical realizations of a given degree sequence is a fundamental component in simulation-based measurements of network observables, with applications ranging from epidemics, through social networks to Internet modeling. Existing graph sampling methods were either link-swap based (Markov-Chain Monte Carlo algorithms) or stub-matching based (the Configuration Model). Both types are ill controlled, with unknown mixing times for link-swap methods and uncontrolled rejections for the Configuration Model. We have solved this problem via several contributions to both the MCMC swap-based sampling and stub-matching methods.

In the stub-matching approach we adopted an entirely different, but rigorous method, by providing a direct construction algorithm, which generates the graph samples with calculable weights. These weights can then be used to account for biases in an exact fashion. For undirected networks the mathematical theorems at the basis of our algorithm were published in *Journal of Physics A: Mathematical and Theoretical* [18]. Due to its important and fundamental nature, after peer review, the journal decided to publish our result as a Fast Track Communication (FTC). According to the journal, FTC's are "outstanding short papers reporting new and timely developments". The algorithm we subsequently developed was published in [19]. For directed graphs (directed graphical degree sequences) the mathematical theorems we developed for our algorithm were published in [20]. The algorithm we developed subsequently was published in [21]. Our algorithms in both cases generates graphical realizations of a degree sequence in polynomial time and statistically independently from one another. Our algorithms do not create the graphs uniformly, but with known weights. These weights then can be used to correct for sampling biases within an importance sampling scheme and generate in the end the graph as if they were sampled uniformly at random. In contrast with previous methods, our algorithms do not require back-tracking or rejections. We have also shown that for large networks, and for degree sequences admitting many realizations, the sample weights are expected to have a lognormal distribution. This is shown in Fig 35 on examples, where we applied our algorithm to generate networks with degree sequences that were drawn from power-law distributions.

The second general approach to graph sampling is based on Markov Chain Monte Carlo (MCMC) using edge-swaps. The algorithm itself is much much simpler than those by direct construction, however, its mixing time is one of the hardest open math problems today, and since it is unsolved, MCMC graph sampling simulations are ill-controlled. We have to start from an element of the ensemble of graphs with a given degree sequence, and then we choose at random independent pairs of edges and swap the ends. This operation preserves the degree sequence but it changes the graph. The mixing time, closely related to the relaxation time is the typical number of swaps we need to do in order to approach the uniform stationary distribution of the Markov chain within a prescribed distance. This distance is measure usually as the largest variation distance and can be related to the spectral gap of the Markov transition matrix. With the exception of special classes of degree sequences (usually regular degree sequences) so far no-one has been able to show that the MCMC method mixes fast, i.e., the samples generated loose the memory of the starting point within a polynomial number of steps. We have recently added to the set of fast mixing proofs for another class of degree sequences, more precisely for graph sampling constrained by joint degree matrices (JDMs). These are actually more complex constraints than just degree sequence based.

Joint-degree matrix based sampling of graphs

A joint-degree matrix specifies the number of connections between nodes of given degrees, for all possible degree values. Given a JDM it uniquely specifies the degree sequence, however, a degree sequence admits many JDMs. Modeling of networks based on JDMs allows us to fully control degree-degree correlations, or the level of assortativity. For example, social networks are characterized by strong degree assortativity, whereas most other networks (technical/infrastructure or biological) they are characterized by negative degree correlations. Having efficient algorithms to construct and/or sample networks with given JDMs allows us to construct better models based on data collected from real-world networks. These results apply not only to networks of immediate interest to DTRA, but to network science in general. However, JDM based construction and sampling problems are much harder to tackle, as they require the development of several pure mathematical theorems and their proofs. We managed to bring results in this area as well and published two main papers on this topic. In one of the papers we developed a novel proof technique for fast mixing by Markov Chain Monte Carlo edge-swaps over balanced realizations of graphs with given JDM, published in Siam Journal of Discrete Mathematics [22]. This methodology combined with novel results on graph decompositions into split graphs has recently allowed us to produce a proof for fast mixing for exponentially large classes of novel uni-partite and bipartite degree sequences (manuscript under preparation).

In parallel, we have also developed a direct construction based algorithm for sampling graphs with given JDM, i.e. with prescribed degree correlations. This work was published recently in [23].

The case for more general constraints: How random are complex networks?

We have also undertaken an extended analysis of the levels of randomness found in real-world complex networks, which shows that for most cases, our methods of network modeling based on degrees and degree correlations is sufficient to capture the nonrandom of the network. In particular, one reads off from the data network all the degree correlations within subgraphs of order *at most* d (dk-series) see Fig 36, then randomize the edges of the data network so as to keep all these series (up to d) intact. The original and the randomized data network are then compared. Usually, low-order correlations ($d = 2 - 2.5$) are sufficient to capture all significant network properties in real world networks. This work has been published recently in the journal Nature Communications [24].

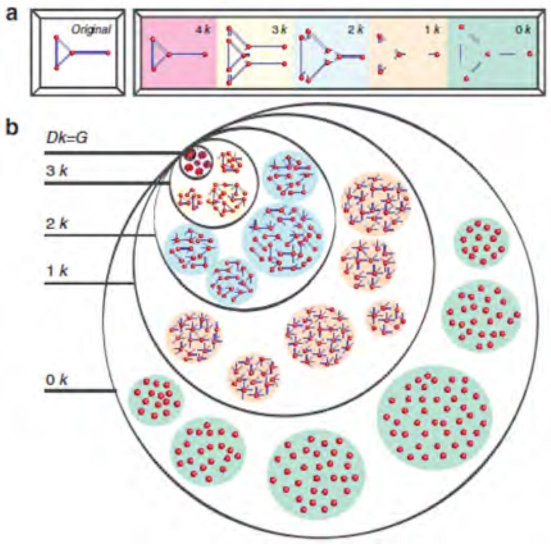


Fig. 36: *dk*-series. 1*k*-series is the sequence of degrees; *dk*-series specifies the sequence of joint degrees of small subgraphs on *d* nodes. Published recently in *Nature Comm.* [24].

9) Network Discovery by generalized random walks

The principal goal here was to develop efficient algorithms that can assess the damages in the shortest time possible in a network. Assuming localized damages in the infrastructure due to a G-WMD one is tasked to find efficiently the part of the network that has been destroyed or incapacitated. In order to find and assess these parts, one needs to interact with the network or send out walker swarms to walk and read/discover the network (software or physical). We have studied the assessment problem under the more general case of network discovery when the whole network is unknown, and focused on the network crawling method. In particular, for efficient explorations of large-scale networks, we have investigated properties of network discovery by walkers that follow edges with some transition probability, in the most general setting of arbitrary but stationary transition probabilities (STP). Let $p_n(s'|s)$ be the probability that a walker, who after n steps is currently at node s, will take the next step onto neighbor s' of s. These single-step transition probabilities are stationary if they are independent of n, i.e., $p_n(s'|s) = p(s'|s)$. We have derived exact expressions for the average number of discovered nodes $\langle S_n \rangle$ and edges $\langle X_n \rangle$ after n steps (more precisely for their generating functions), for arbitrary graphs and arbitrary STP transition probabilities (STP walkers). Let

$$S(\xi) = \sum_{n=0}^{\infty} \xi^n \langle S_n \rangle \quad \text{and} \quad X(\xi) = \sum_{n=0}^{\infty} \xi^n \langle X_n \rangle \tag{17}$$

be the corresponding generating functions and let $P(s'|s;\xi) = \sum_{n=0}^{\infty} \xi^n P(s'|s;n)$ be the generating function for the site occupation probabilities $P_n(s'|s)$. Then:

$$S(\xi) = \frac{1}{1-\xi} \sum W(s;\xi) P(s|s_0;\xi) \tag{18}$$

and

$$X(\xi) = \frac{\xi}{1-\xi} \sum_{s \in V} \bar{W}(s;\xi) P(s|s_0;\xi) \tag{19}$$

with

$$W(s;\xi) = b^{-1} \quad \text{and} \quad \bar{W}(s;\xi) = \sum_{s' \in V} \alpha \, \frac{1 + (d-c)\beta\xi}{1 + (a\alpha + d\beta)\xi + (ad - bc)\alpha\beta\xi^2} \tag{20}$$

where $b = b(s;\xi) = P(s|s;\xi)$, $c = P(s'|s';\xi)$ are return probability generating functions, $a = P(s|s';\xi)$, $d = P(s'|s;\xi)$ and $\alpha = p(s'|s)$, $\beta = p(s|s')$ are the STPs. These exact expressions, valid for any STP walk and any connected graph allows us to study the properties of network discovery as function of time (the corresponding probabilities can be obtained from the generating functions via Cauchy's formula or via several asymptotic methods). In particular, using these expressions we have shown that for STP walkers both edge and node discovery follows the same scaling law on large networks, independently on the form of the stationary transition probabilities. Hence, the discovered or the visited part of the network will be sparse (the number of discovered edges scaling linearly with that of the discovered nodes), presenting a *strongly skewed* structure compared to the underlying network's (the true structure). Only after a crossover time of $\mathcal{O}(N)$ (where N is the number of nodes) will the edges become increasingly discovered, which in the case of large networks means unfeasibly large wait times. Thus via rigorous proofs we eliminated STP walks as an efficient methodology for faithful edge discovery on large networks. One can therefore easily miss damaged edges via crawling using time-independent transition probabilities. Our results thus rigorously show that efficient/faithful discovery can only be done with adaptive walkers, whom use time/history dependent information for their transition probabilities (ATP). Visiting history information can be thought of as "pheromone" trails on the network, which the walker uses through its rules for stepping onto the next site. There is a plethora of possible rules using past history, however, to keep memory requirements low (bounded) on a walker, the desirable rules are the ones that only use information from the local neighborhood of the walker. In this vein, we have introduced a simple adaptive walk, the Edge Explorer Model, which is greedily biased towards already visited regions within a 2-step neighborhood. We have shown that on dense graphs the EEM performs near optimally or optimally (Fig 37). These results have been published in [25].

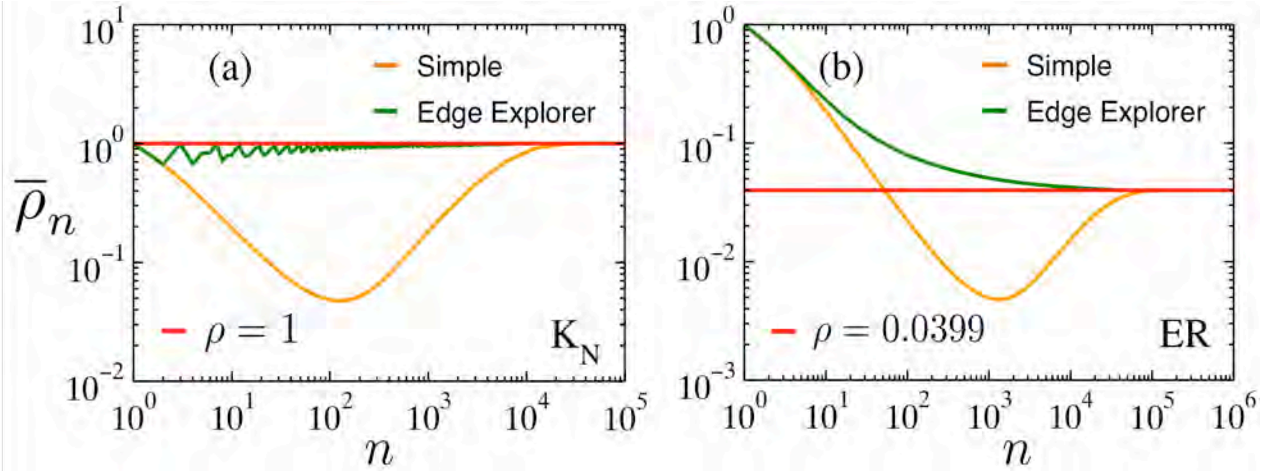


Fig 37 : The Edge Explorer Walker on a) complete graph (worst case) and Erdos-Renyi random graph b). The y-axis shows the density of the discovered part of the network; the red is the true density of the underlying network. The yellow curve shows the measured/perceived density of the discovered network as function of time by a simple random walker (thus STP type) and the orange green shows the the same quantity for the Edge Explorer Walker (ATP type).

10) Translational results

Many of the methods and theories developed in this project are general and can benefit the larger class of network science research problems. Many of our methods have been and are being applied by other authors. However, we have also made use of these approaches on some of our other projects, e.g. in [26]. In particular, we have used the network modeling techniques such as the MaxEnt method, and the betweenness centrality results on brain neuronal networks. The data for these works have been generated by a neuroanatomy group from Lyon, France, using retrograde tracer experiments in the macaque brain and also in the mouse brain. Our modeling approach has revealed a previously unsuspected organization of the interareal cortical network. Our network model, called the Exponential Distance Rule model (EDR) is able to capture (without fitting parameters!) the structural characteristics of the real cortical network in the macaque, and most recently in the mouse as well and predicys well new experimental findings. These findings have been published in several publications, in top neuroscience journals, such as Neuron, Cerebral Cortex and PNAS, all acknowledging DTRA support. These are: [27], [28], [29], [30], [31].

B) Bio-pathogen WMD (E-WMD) related work

1) Multiscale interactions in reaction-diffusion models

During the life of the project we completed the analysis and integration of a variety of multiscale mobility networks worldwide. This has allowed us to construct the first synthetic multiscale mobility networks describing the daily human mobility at the worldwide scale. The extensive analysis of these networks has allowed us to draw a general gravity law for commuting flows that reproduces commuting patterns worldwide. This law is statistically stable across the world due to the globally homogeneous procedure applied to build the subpopulations around transportation hubs. The gravity law allows us to work with two different worldwide commuting networks. An entirely synthetic one, generated using the gravity law fitted to the empirical data, and one integrating the empirical data. The worldwide mobility network has been used to study the impact of the network multiscale features in the dynamical behavior and patterns of spreading and diffusion process. In order to deal with the integration of computing intensive multiscale equations we have developed a mechanistic approach coupling different elements of the mobility network through effective interactions accounting for the mobility of individuals.

These results have been published on the Proceedings of the National Academy [3], and have been integrated in a general computational platform that can be used to study diffusion processes at the worldwide scale. We specify the definition and integration of the different data layers composing the model and the computational platform in a BMC Infectious Disease publication [32]. The construction of the mobility network and the derivation of the stochastic mobility equations among different subpopulations are described in detail as well. We illustrate the time-scale separation technique that allows for the integration of the mobility processes occurring on small time scales as effective coupling terms. This method reduces the computational cost by simulating in an explicit way only mobility processes occurring on the long time scales.

In order to assess the level of realism offered by the multiscale approach we have also undertaken a major effort to perform a side by side comparison of the developed computational model with an extremely detailed agent based model [33]. Indeed, agent-based models provide a very rich data scenario but the computational cost and most importantly the need for very detailed input data has limited its use to country level. On the opposite side, the multiscale model is scalable and can be conveniently used to provide contagion world-wide scenarios and patterns with thousands of stochastic realizations. In this perspective, it is clearly important to assess the level of agreement that the two different approaches can provide on the quantities accessible in both cases and the respective data needed and computational costs associated. The results obtained in our study show that both approaches provide contagion patterns that are in very good agreement at the granularity levels accessible by both approaches. This is an extremely encouraging result that paves the way to the construction of hybrid models combining the agent-based and the multiscale approaches according to the available data and computational resources.

The availability of the computational platform integrating multiscale networks has led to the development of interactive network visualization tools of non-local damage spreading in complex-multiscale networks. In particular, we have been working on the analysis and network visualization of the epidemic invasion tree of epidemics in geographically structured populations. Humans are the carrier of the pathogen, and the large-scale travel and commuting patterns that govern the mobility of modern societies are the substrate over which epidemics and pandemics travel. Microsimulations of epidemic outbreaks can provide information at very detailed spatial resolutions and down to the level of single individuals. In addition, computational implementations explicitly account for stochasticity, allowing the study of multiple realizations of epidemics with the same parameters' distribution. While on one hand those capabilities represent the richness of microsimulations methods, on the other hand they face us with a huge amount of information that requires the use of specific data reduction methods and visual analytics. In the study of the global spreading of epidemics, it is possible to develop models that keep track of single individuals flying on specific airline connections that carry the infections from one area of the world to another (generally urban or census areas). This produces a considerable amount of data in each single microsimulation of the epidemic process; data which are then multiplied by the number of stochastic realizations needed to extract meaningful statistical patterns.

For this reason, one convenient data reduction technique consists in the mapping at the global level of the geographical infection trees. These networks identify in each microsimulation the "infector" city of each "infected" city. In other words, every time an individual carries the disease from a city i to city j, we draw a link that shows that the infection propagated along that connection. On its turn, each "infectee" city can be the 'infector" of another city, giving rise to a tree that characterizes the spatial spreading of the disease at the city level, as measured from the movement of single individuals. By repeating the microsimulation process, each city j can have a single infector source i, or several possible infector sources. The average of p_{ij} over all realizations will give us the resulting probability of city i being the infector of city j. To obtain the infection tree from a set of simulations sharing the same source of the outbreak, one can use the Directed Minimum Spanning Tree of the corresponding network using a Chiu-Liu/Edmonds algorithm, where the cost of each edge wij is given by the probability 1-p_{ij}. This procedure will result in a directed spanning tree where every node j has no more than one incoming edge, corresponding to the most likely infection path that leads to the outbreak at each infected node. In Fig 38 we show the infection tree of a numerical pandemic obtained with the Global Epidemic and Mobility (GLEAM) computational model.

The infection trees provide a cartography of the evolution of pandemics starting from different seeds as soon as we know the basic parameters governing the risk of infection of individuals and duration of the disease. The infection tree also lends itself to further analysis in which it is possible to evaluate the role of each node or group of nodes in the spread as well as shed light on the global circulation of seasonal influenza.

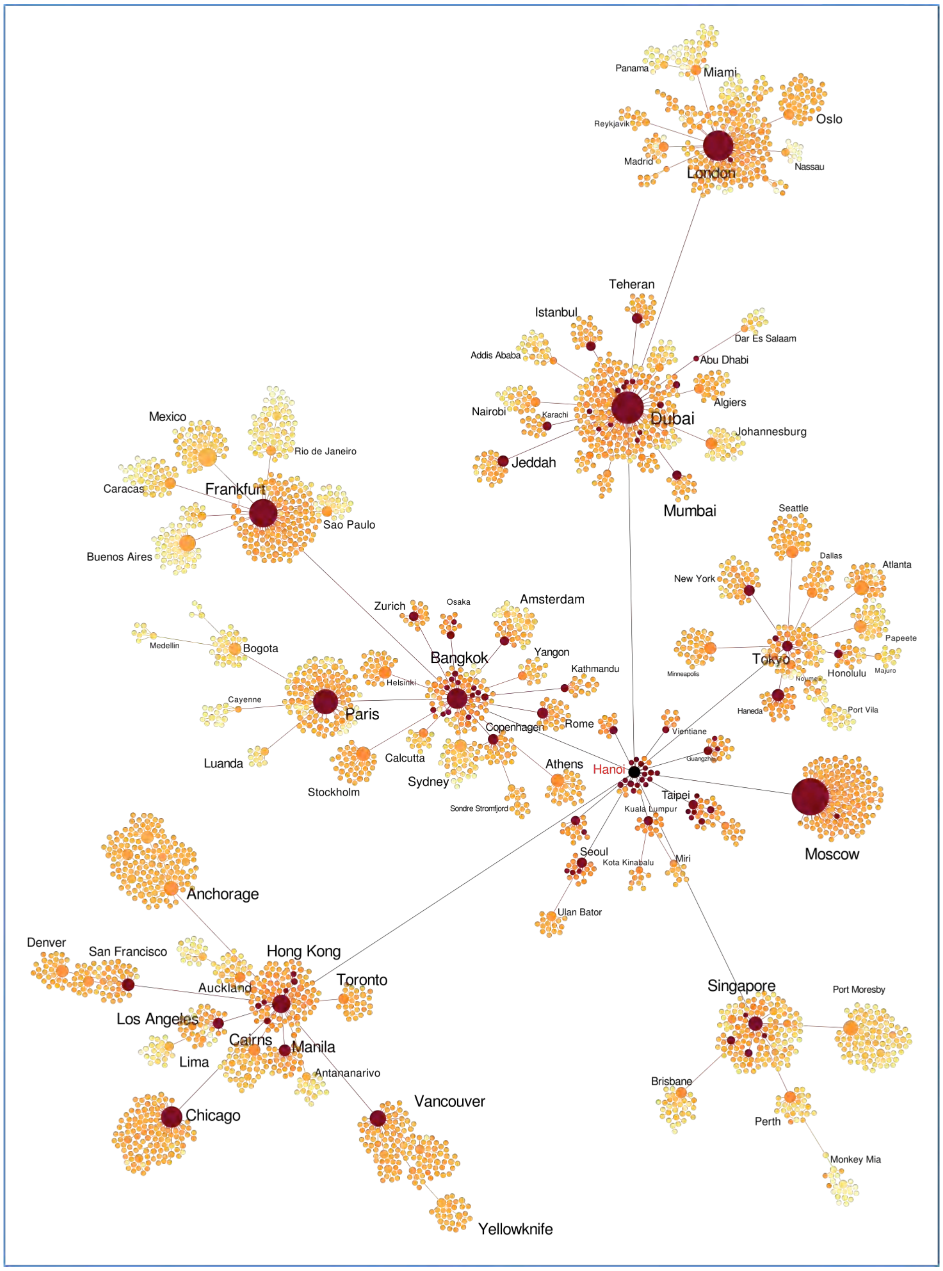


Fig 38 : Network representation of the infection tree. The origin of the outbreak is the city of Hanoi, Vietnam, colored in black. The color code for the nodes goes from dark red, meaning early arrival of the disease, to light yellow for wider time gap between the outbreak in Hanoi and the arrival on the corresponding node. The size of each node is proportional to the degree. The labels are related to the hub at the corresponding branch.

The use of the infection tree provides a meaningful representation of the epidemic spreading, statistically aggregating the microsimulations results, thus proving this data reduction technique a valuable visual tool in the mapping of the progression of epidemics in the case of scenario analysis and pandemic preparation plans. The invasion tree analysis has been published in the journal "Network Science" [34].

The developed algorithms for multiscale networks have all been integrated in the publicly available computational platform Global Epidemic and Mobility model (GLEAM). This also includes the development of appropriate plug-ins for the analysis of the intentional release of a E-WMD pathogens, along the necessary documentation for their use in the GLEaMviz tool (www.gleamviz.org).

2) Resilience of structured population networks

At the early stage of the project we have focused on the analysis of the onset of specific critical tipping points in the spreading of biological agents and information processes in populations characterized by highly regular mobility networks. Human mobility patterns have been shown to be highly predictable and generally dominated by a set of specific locations and recurrent bi-directional flows across those locations, the typical example being the everyday commuting patterns to the work place of millions of individuals. These recurrent mobility patterns are poorly modeled by the random diffusion processes generally used to study spreading processes. We have developed a framework based on a time-scale separation technique to analyze the behavior of contagion and spreading processes in a network of locations where individuals have memory of their location of origin. We find a phase transition between a regime in which the spreading phenomenon affects a macroscopic fraction of the system and a regime in which only a few locations are affected. For disease and information spreading processes the phase transition critical point defines an invasion threshold that depends also on the mobility parameters, providing guidance on how to control disease and information spreading by constraining mobility processes. We recover the threshold behavior in the analysis of diffusion processes mediated by the real data of human mobility pattern.

In order to develop a model of highly regular mobility patterns we have considered the prime example of commuting flows. In this case we consider a metapopulation system with V distinct subpopulations, each of which has a population size N_i. The V subpopulations form a network in which each subpopulation i is connected to a set of other subpopulations U(i). The edge connecting two subpopulations i and j indicates the presence of a flux of commuters. We assume that individuals in the subpopulation i will visit anyone of the connected subpopulations with a per capita diffusion rate σ. As we aim at modeling commuting processes in which individuals have a memory of their location of origin, displaced individuals return to their original subpopulation with rate t. These parameters thus influence the probability that individuals carrying infection or information will export the contagion process in nearby subpopulations. If the diffusion rate is approaching zero, the probability of contagion of neighboring subpopulations goes to zero as there are no occasions for the carriers of the process to visit those. On the other hand, if the return rate is very high the visiting time of individuals in neighboring populations is so short that they do not have time to spread the contagion in the visited subpopulation. In order to find an explicit expression for the parameters' value that separate the regime in which the contagion process is suppressed from the regime in which the contagion process invades the entire system we have used a time scale separation technique that defines quasi-stationary mixed subpopulations (see Fig. 39). Each mixing subpopulation identifies the number of individuals N_{ij} of the subpopulation i visiting subpopulation j. This approximation allows the writing of a subpopulation branching process for the contagion process at the subpopulation level, yielding an explicit form for the condition that defines the invasion of a macroscopic fraction of the subpopulation network; i.e. the percolation transition of the contagion process. We have performed accurate simulations for the study of the invasion threshold in synthetic graphs and realistic commuting networks in the USA and find that the results are in very good agreement with the analytical predictions. The findings of this part of the activity resulted in high impact publications in the journal Nature Physics and Journal of Theoretical Biology [35], [36]. These publications are opening the path to the inclusion of more complicated mobility or interaction schemes and at the same time provide a general framework that may be used not just as an interpretative framework but a quantitative and predictive framework as well. Understanding the effect of mobility and interaction patterns on the global spread of contagion processes can indeed be used to devise enhanced or suppressed spread by acting on the basic parameters of the system in the appropriate way, which might find

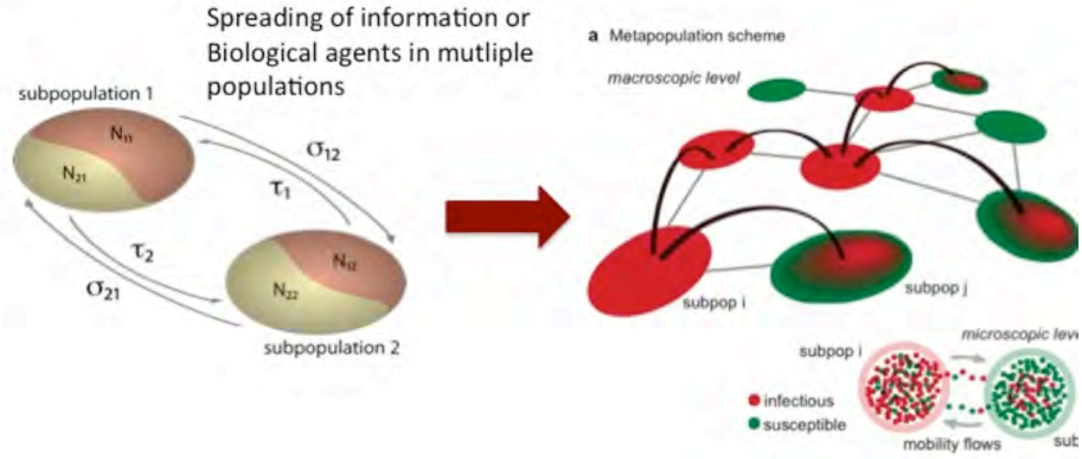


Fig 39 : Modeling scheme for the analysis of the resilience to contagion processes of structured populations coupled by commuting mobility patterns. Left: Coupling of two subpopulations by commuting fluxes defining mixing subpopulations. Right: Invasion process of the contagion process via the mobility of individuals.

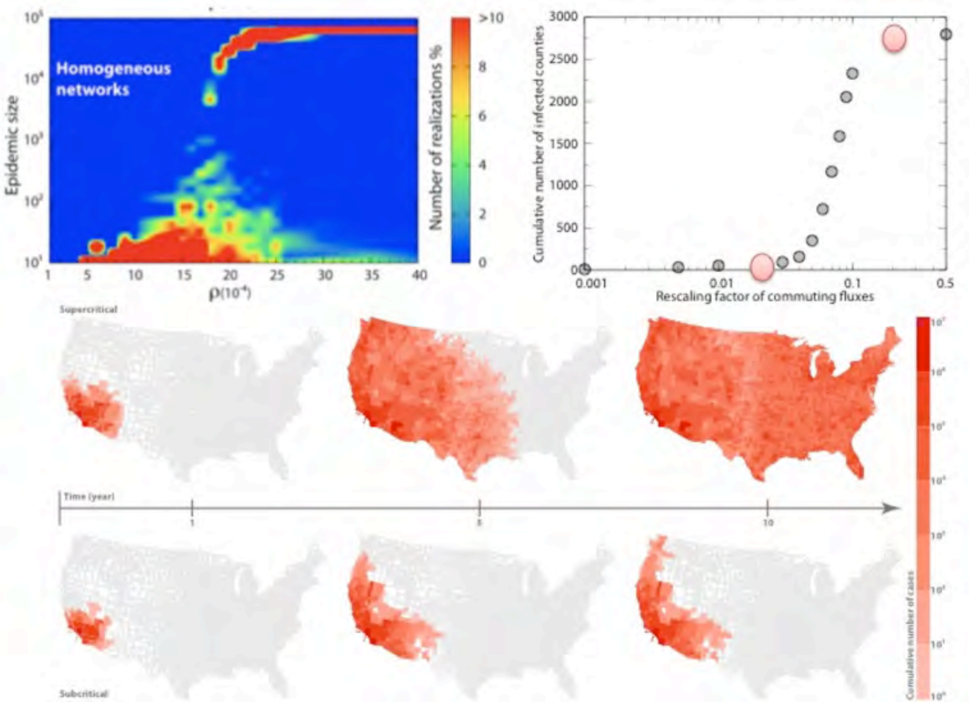


Fig. 40: Analysis of the contagion behavior and the associated invasion threshold. Upper plots: Both plots report the behavior of the number of subpopulations invaded by and susceptible-infected-removed contagion process as a function of the commuting rate. On the left we show the results from numerical simulations in a synthetic network. On the right we report the simulations for the network defined by the commuting fluxes among USA counties. The bottom panel compares the spreading in time of the contagion process in the USA county network below and above the invasion threshold. Below the invasion threshold the contagion process spontaneously dies out while above the threshold it invades almost 100% of the USA.

applications ranging from the protection against emerging infectious diseases to viral marketing. For this reason, the basic algorithm and schemes defined during the research activity have been integrated in a general computational approach to infectious disease modeling. A detailed presentation of the global epidemic and mobility model (GLEAM) including the contribution due to recurrent mobility patterns has been published in the Journal of Computational Science [37].

The geographical structure of population is at interplay also with the social network structure of the population experiencing a biological threat. For this reason, we have also achieved several results concerning the coupling of self-initiated behavioral changes in the progression of contagion phenomena. Behavioral changes vary from simply avoiding social contact with infected individuals and crowded spaces to reducing travel and preventing children from attending school. In all cases we have a modification of the spreading process due to the change of mobility or contact patterns in the population. In general, these behavioral changes may have a considerable impact on epidemic progression such as the reduction in epidemic size and delay of the epidemic peak. We have carried out two different studies aimed at coupling contagion models with social adaptive behavior.

The first study aims at developing a general mathematical framework describing the different mechanisms that consider the spread of awareness of a disease as an additional contagion process. Three mechanisms were proposed. In the first, basic model the social distancing effects and behavioral changes are only related to the fraction of infected individuals in the population. In the second we modeled the spread of awareness considering only the absolute number of infected individuals as might happen in the case that the information the individuals rely on is mostly due to mass media reporting about the global situation. Finally, in the third model we added the possibility that susceptible people will initiate behavioral changes by interacting with individuals who have already adopted a behavioral state dominated by the fear of being infected. This apparently simple interaction allows for the self-reinforcement of fear. We have found that these simple models exhibit a very interesting and rich spectrum of dynamical behaviors. We have found a range of parameters with multiple peaks in the incidence curve and others in which a disease-free equilibrium is present where the population acquires a memory of the behavioral changes induced by the epidemic outbreak. This memory is contained in a stationary (endemic) prevalence of individuals with self-induced behavioral changes. Finally, a discontinuous transition in the number of infected individuals at the end of the epidemic is observed as a function of the transmissibility of fear of the disease contagion. The results of this study have been collected in paper published on the interdisciplinary journal [38].

The second line of work concerned the coupling of behavioral changes of the population with the large-scale network describing the mobility of people across location in the USA. This work, that builds also on the above general results concerning the coupling of contagion processes and social adaptation, introduce a simple mechanism that provides individuals with the propensity to avoid locations affected by contagion processes. Although the aim of such a self-initiated behavior is to prevent an individual's exposure to the pathogen, it may lead to the unanticipated effect of facilitating its spread to new locations. The results presented in this paper underline the importance of the proper consideration of self-initiated behavioral responses to the spreading of an emerging infectious disease including a possible pathogen due to a biological WMD attack. The results of this study are collected in a paper published in [39].

We have also studied the effect of population structure when multiple infectious agents circulate in a given population of hosts, and they interact for the exploitation of susceptible hosts aimed at pathogen survival and maintenance. The dynamic of infectious diseases has been traditionally studied focusing on single pathogens one at a time, increasing attention is currently being devoted to the interactions among multiple infectious agents. Interaction mechanisms can indeed alter the pathogen ecology and have important evolutionary, immunological and epidemiological implications. We used mathematical and computational model for disease transmission between hosts and for the mobility of hosts in a structured metapopulation network to study the competition between two pathogens providing each other full cross-immunity after infection. Depending on the rate of mobility of hosts, competition results in the dominance of either one of the pathogens at the spatial level – though the two infectious agents are characterized by the same invasion potential at the single population scale – or co-circulation of both. We have also explored explore systematically the role of epidemiological and immunological (i.e. reproductive numbers of the two pathogens and cross-immunity) and ecological parameters (spatial distribution of the hosts and mobility) in defining the co-circulation dynamics of competing pathogens. The competition dynamics is reconstructed through extensive numerical simulations of a stochastic mechanistic model i.e. a multispecies reaction-diffusion model in a complex metapopulation network (see Fig. 41).

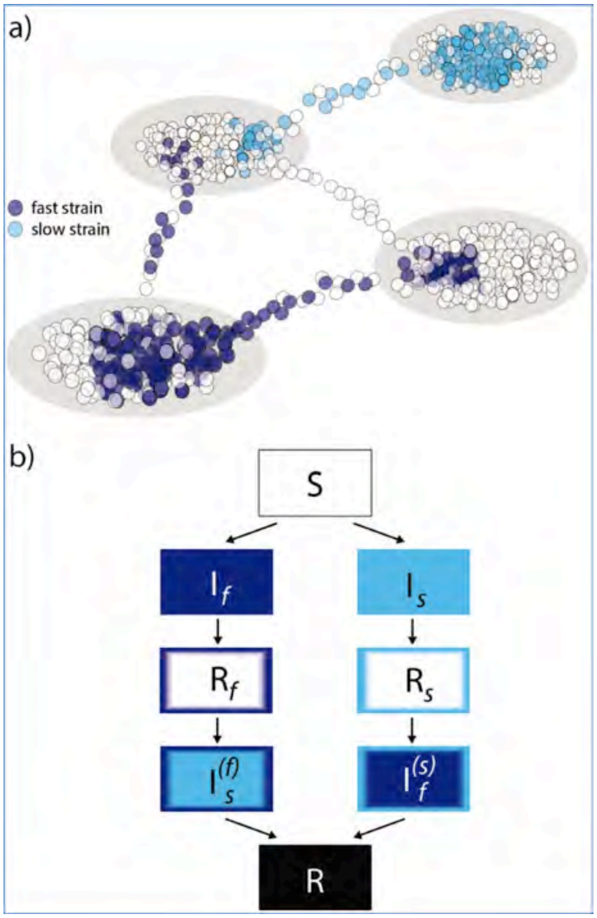


Fig 41: (*a*) Scheme of the metapopulation structure in patches and links representing mobility. (*b*) Individual state of the compartmental model of the two-strain infection.

We provided an extensive numerical characterization of the interaction dynamics by varying the degree of cross-immunity and the difference in the epidemiological traits of the two pathogens (basic reproductive numbers and infectious periods). Depending on the relation between the pathogens' traits, mobility can play a determinant role or be non-influential for the outcome of the competition. In the space of epidemiological parameters, there exists a region for which lowering the traveling probability induces a cross over from the fast strain dominance to the slow strain one. This behavior is determined by the trade-off between epidemic growth and potential for spread at the spatial level, the former being an advantage in a well mixed population while the second being relevant in a sparse environment with intermediate or low mobility coupling. Our study characterizes the epidemiological conditions under which hosts' traveling behavior is a determinant ingredient in the competition dynamics. Reducing the degree of cross-immunity determines a rapid transition between the picture described here to a situation of no competition in the spatial propagation. This transition behavior can be framed within a herd immunity paradigm, where the more rapid pathogen acts as a vaccine in the spreading dynamics of the other one.

The modeling framework here introduced allows us to account for important features characterizing host mobility patterns in a realistic way. Despite the complexity of the dynamics simulated by the mechanistic model, the analytical formulation of the global invasion potential allows for simple analytical considerations able to shed light on the behavior observed in numerical simulations. The network formalism is an important ingredient of the model and represents an element of novelty with respect to previous studies on disease ecology where the space is introduced by placing individuals on a regular lattice or by assuming a metapopulation with two levels of mixing (high mixing within patches and low mixing between patches). The results concerning structured populations and competing pathogens have been summarized in papers [40], [41].

3) Mitigation and containment of "Secondary" Epidemic (E-WMD) threats

The escalation of terrorist activities in the last decade has spawned a growth of interest in modeling the course of a deliberate introduction of a highly pathogenic agent such as smallpox under different scenarios. Studies have been mostly focusing on strategies to contain local outbreaks after their detection, showing that timely coordinated intervention with vaccination, contact isolation and tracing are able to halt ongoing transmission. All studies however do not consider the possibility that the present human mobility volume could escalate to an international dimension the smallpox outbreak even before an official warning could be issued. For this reason we have studied simulate the international spreading of smallpox release in the time preceding the declaration of the contagious disease emergency. We show that in wide range of plausible values for the key epidemiological parameters and release scenarios, index cases and potential outbreaks can occur in several countries in different continents before an international emergency is declared. This has two major implications: i) bioterrorism targeting only a specific country is instead going to have global effects; ii) the attack may eventually trigger outbreaks in countries where the deployment of effective containment measure and policies is unlikely because of the lack of health infrastructure and medical resources. In particular, we have focused on the level of threat for USA in the case of a bioterrorist attack outside its border, namely a European country such as UK or France.

What we are interested in is the quantitative assessment of the risk of the internationalization of the outbreak and the number of countries involved. Building on previous works of this project (especially the computational platform published in [32], we have developed a large-scale structured metapopulation model that accurately describe the worldwide spread of smallpox during the initial period of time between the occurrence of the intentional virus release and the issuing of an international emergency. Clearly, this risk assessment depends on the time elapsed from the biological attack. The point of interest is therefore the time at which the international community is able to issue a worldwide alert and starts implementing containment and mitigation policies. This includes the ability of an effective contact tracing, the deployment of vaccine stockpiles for ring vaccination, travel restrictions, etc. Theoretically, the earliest time at which the detection can occur is when the first person in the rash stage is correctly diagnosed.

Country	Initial detection		After 1 week		After 3 weeks	
	Outbreak probability (%)	Conditional number of cases	Outbreak probability (%)	Conditional number of cases	Outbreak probability (%)	Conditional number of cases
UK	100	28 [18-40]	100	68 [43-102]	100	239 [152-355]
USA	33	1 [1-5]	56	2 [1-11]	96	9 [1-41]
Germany	18	1 [1-4]	34	1 [1-9]	82	4 [1-30]
France	16	1 [1-4]	31	1 [1-10]	81	4 [1-30]
Italy	15	1 [1-4]	29	1 [1-9]	76	3 [1-28]
Spain	15	1 [1-5]	29	1 [1-10]	76	3 [1-31]
Ireland	13	1 [1-4]	25	1 [1-9]	70	3 [1-25]
Netherlands	9	1 [1-4]	18	1 [1-8]	59	2 [1-22]
Switzerland	9	1 [1-3]	17	1 [1-8]	55	2 [1-20]
Canada	6	1 [1-3]	11	1 [1-9]	40	2 [1-24]
Belgium	5	1 [1-4]	10	1 [1-8]	38	2 [1-22]

Table 1

Due to the current rarity of the disease and the difficult diagnostic, it is likely that an alert would only be possible after several civilian cases had already occurred.

In our study we assume that 4 individuals are required to be hospitalized before successful detection is feasible and an alert issued and perform a sensitivity analysis in the range of 2 to 8 individuals. We are therefore measuring the distribution of the number of countries affected and the specific risk posed for each individual country at detection time. In addition, from the successful detection of the smallpox outbreak to the implementation of effective contact tracing and a worldwide coordinated response, including the deployment of vaccine, most of the experts consider a time window ranging from 1 to 4 weeks. For this reason, we report data also for the four consecutive weeks proceeding the outbreak detection.

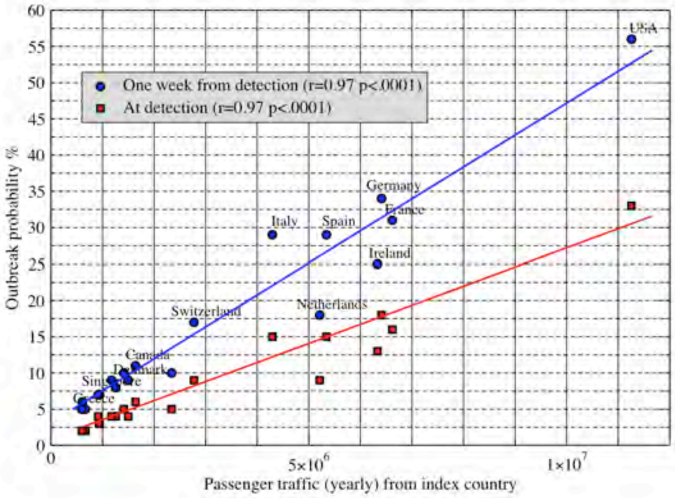


Fig 42: Global spread through the human worldwide mobility network of a highly pathogenic agent (smallpox) after a localized intentional release. Correlation between the probability of experiencing an outbreak (secondary WMD event) in a given country and the traffic with the initially targeted country. In the present figure the initial target country is UK (city of London).

Our findings are summarized in the following table where we report the top 20 countries at risk providing the probability of infection and the expected number of cases.

Not only is the risk never geographically restricted, even at the time of detection or just a week later when the first interventions might be feasible (see Table 1 and Fig 42), it also increases rapidly both in terms of exposure size and probability. The reason for such a rapid global spread is due to the fact that London serves as a major connecting point between the various continents, with daily connections to Europe, Africa, Asia and the Americas. However, other major hubs in the USA or Europe produce the same results as shown in the published paper [42].

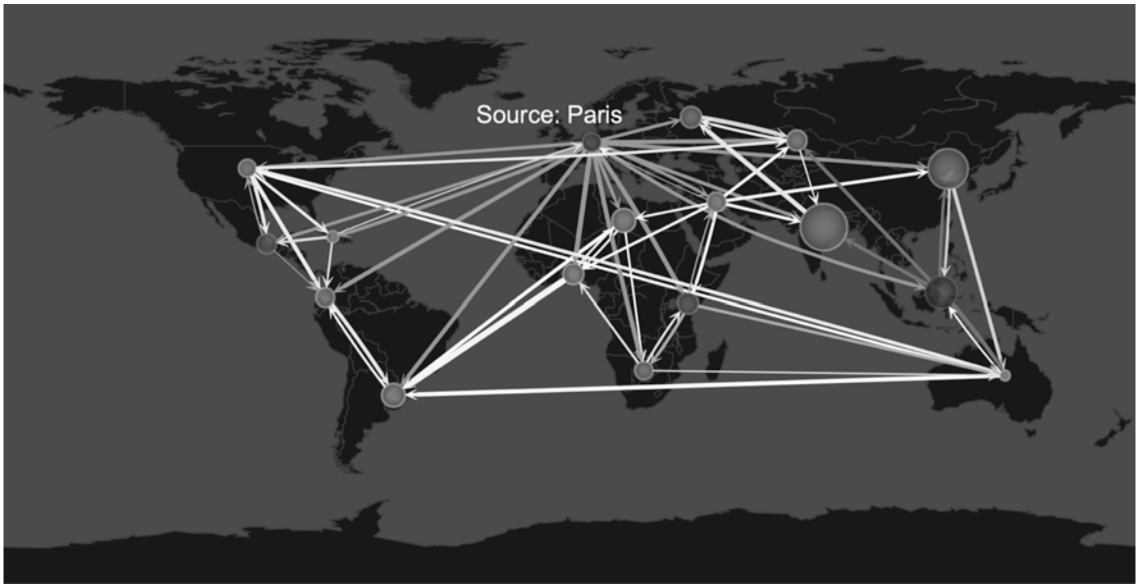


Fig 43: Global spread through the human worldwide mobility network of a highly pathogenic agent (smallpox) after a localized intentional release. Invasion tree mapped at the level of the WHO surveillance regions. In the present figure the initial target country is France (city of Paris). The tree shows the circulation of the infectious cases and the corresponding seeding of local outbreaks.

Our research shows that biological targeted attacks on a single country or city can result in a global threat. For instance, an attack in the UK or any European countries is very likely resulting in a WMD threat on the USA. For instance, while previous studies have pointed out that the timely response to a smallpox outbreak is very likely to be successful, those accounts have only considered the case of single countries with very effective health systems and medical resources. The present work indicates that a deliberate smallpox release is very likely to assume an international dimension even before the outbreak is identified.

In order to quantify in more detail, the impact of human mobility network in the analysis of secondary E-WMD threat we have studied the correlation between the probability of any given country of experiencing the importation of infectious individuals and the traffic with the initial targeted country (see Fig.42). We find a strong correlation that however indicates that traffic reductions are not going to drastically reduce to zero the importation probability unless implemented with 90% efficacy. As the overall traffic reduction is not a viable option, we have been focusing on optimal mobility control strategies and/or network restructuring that lower the non-local spread across subpopulations helping the mitigation and containment of secondary E-WMD threats.

We have been working with the Global and epidemic Mobility model developed in the previous years to draw scenarios from and define risk assessment metrics to threats posed by E-WMD in a wide spectrum of initial conditions and localized releases. The goal is the understanding of the long-range spreading pathways defined by the human mobility network (see Fig. 44). We have proceeded by using an analysis of the geographic invasion tree capturing both the hierarchy and the most probable transmission routes of the released pathogen released in an E-WMD event.

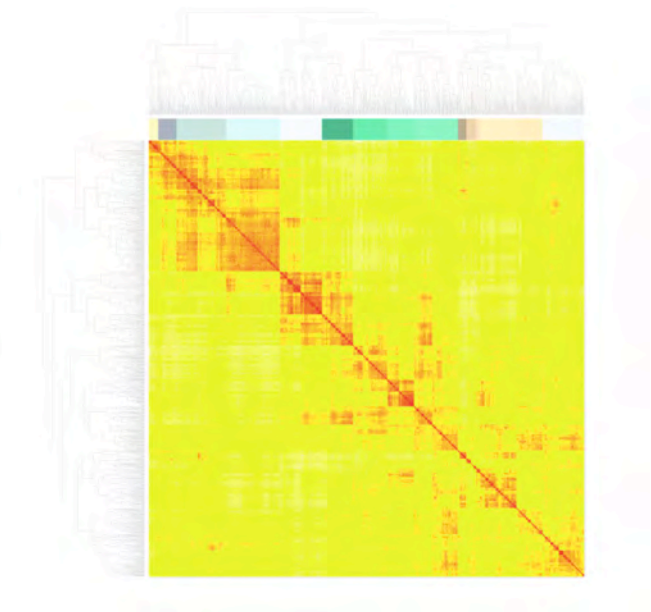


Fig 44: Global spread through the human worldwide mobility network of a highly pathogenic agent (smallpox) after a localized intentional release. Correlation matrix and clustering of the similarity between 4,000 WMD urban areas targets spread across the world. The clustering allows the identification of urban areas generating similar invasion trees worldwide for the pathogenic agent.

The invasion network of the E-WMD event is generated by performing several simulations (runs) for the same initial target k, and calculating for every city (or country pair i and j) the conditional probability p_{ij} that city i export infectious individuals to city j when the E-WMD event starts at source k. The invasion network is then constructed by taking this value as the weight for the edge between cities i and j. An invasion tree can be generated by considering the minimum spanning tree defined by an appropriate metric defined by the weights of the invasion network. With the invasion tree in hand, we can try to obtain the collection of sources that generates a similar worldwide flow. This would allow to group location according to the similarity of the secondary events pattern and define control regions for the mitigation and control of the E-WMD event. In order to establish the similarity between the invasion tree of different initial E-WMD target we have defined several distance metrics characterizing the similarity of networks. Namely we have considered Minkowski, Hellinger, Jaccard and Tversky similarity measures. For all these measures we have defined the similarity matrix M among 4,000 possible E-WMD targets across the world (Fig 44). This matrix allows the clustering of E-WMD targets in groups defining very similar spreading pattern and surveillance areas (see Fig 45). On the basis of the obtained clustering we are working on:

1. General classification of urban areas according to quantitative risk assessment metrics of secondary E-WMD threats.
2. Optimal mobility control strategies informed by the invasion tree structure.

The results presented here have been summarized in a paper published in Scientific Reports [42] In parallel in order to validate the effectiveness of the modeling strategy in the analysis of the early global spread of highly pathogenic viruses, we have performed the study of the early spreading of the H1N1 pandemic, recovering results in simulating case importation, travel restriction efficacy and the global timeline of the pandemic in very good agreement with empirical data. These results have been published in PLoS One [43].
A specific tutorial, reported among the publications in this report, has been prepared to provide guidance in the use of the computational platform GLEAMViz in the study of smallpox like E-WMD events.

Achieving the containment of pandemic potential pathogens is also strongly dependent on the natural history of the disease. For instance, the severe acute respiratory syndrome (SARS) epidemic has been successfully contained by public health interventions that succeeded in reducing the transmissions among individuals. This has not been the case of the H1N1 2009 influenza pandemic that has assumed global proportion in a few months. A key feature of the disease that determines its controllability is the amount of asymptomatic transmission events; i.e. the fraction of infections due to infectious but asymptomatic individuals. The key difference

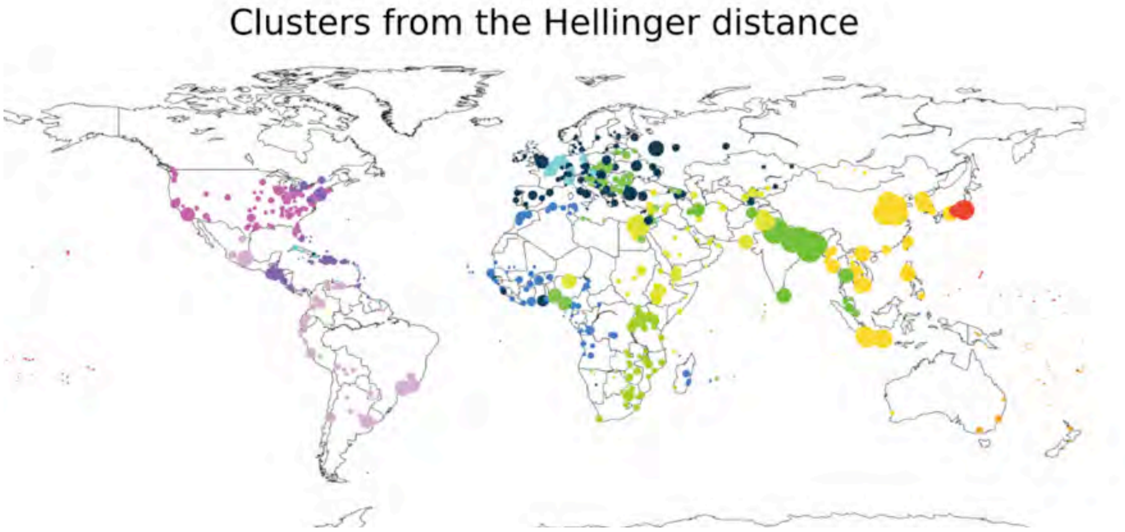


Fig 45: Global spread through the human worldwide mobility network of a highly pathogenic agent (smallpox) after a localized intentional release. WMD targets clustering. Colors identifies groups of urban areas that give rise to very similar invasion tree for the pathogenic agent. Each group defines different containment strategies based on the restructuring of the mobility network.

between these two cases is the amount of asymptomatic transmission events. The larger the number of non-detectable transmission and the smaller is the likelihood of containment. For potential pandemic pathogen, however, the spatial structure and mobility of the population are crucial features that contribute to the global spread of the infectious disease. It has been recently shown that along the local epidemic threshold, which depends only on the disease parameters and is responsible for the epidemic outbreak within each subpopulation, structured populations may exhibit a global invasion threshold that determines whether the system is globally invaded by the contagion process. This novel threshold depends on the mobility rates and patterns of individuals [35], [36] and cannot be uncovered by continuous deterministic models as it is related to the stochastic nature of the mobility and contagion processes. For potential pandemic pathogens, it is therefore extremely relevant to quantify the success of control measures in averting global spreading and how the large-scale spatial controllability on the outbreak depends on the proportion of non-detectable transmissions.

We have developed a metapopulation model in which subpopulations are connected through heterogeneous fluxes of individuals. In each sub-population the epidemic dynamics is characterized by an influenza-like-illness compartmental model with tunable amount of non-detectable infection processes. At the single population level, we can recover the condition for the containment of the local outbreak by means of the isolation of symptomatic individuals as a function of the asymptomatic infections and the inherent transmissibility of the pathogenic agent (as measured by the basic reproductive number R0. We are also able to provide analytical expression for the global controllability of the outbreak defined as the condition that although the disease may spread in a single population, it is not able to spread across multiple populations of the metapopulation system. We interestingly observe that the local and global controllability are defined by different conditions that defines a region of parameters where while the global spreading is controllable, local control at the level of single population is not. This region defines a new window of possibility for the containment at least of the large-scale spreading of the infectious disease. We find that the global controllability region decreases with the increasing of the proportion of asymptomatic transmissions. We investigate the effect of reduction of individual's mobility in defining the extent of the global controllability region and find that only severe mobility restrictions are sorting out considerable effects.

We have continued our work on the determinants of the global controllability of E-WMD outbreaks by looking at optimal containment strategies. In the case of international Potential Pandemic Pathogens release the time needed to develop effective vaccines or deploy large scale pharmaceutical interventions spans weeks or months and control measures such as isolation, contact tracing and quarantine are the first line of defense in the case of new emerging infectious diseases. Achieving the containment of pandemic potential pathogens is however strongly dependent on the natural history of the disease. During period of performance of the grant we have studied strategies for the containment of bio-WMD events relaxing the strict condition of the local containment at the source. We have been looking for a definition of controllability characterized by a region of parameters where, even though at the single population level the outbreak is not controllable, it is possible to control the global spreading of the bio-WMD event. In order to do that we have focused on the containment of local outbreaks by means of the isolation of symptomatic individuals as a function of the rate asymptomatic infections θ and the inherent transmissibility of the pathogenic agent (as measured by the basic reproductive number R0). In particular, we have considered those results in the context of very detailed agent based models that considers the network of contacts among individuals explicitly by creating a synthetic population that specifically considers household structure, workplace and school settings. This modeling approach allows explicit integration of different socio-demographic features via the construction of different synthetic populations. We considered synthetic populations for Mexico, Italy and the Kingdom of Saudi Arabia (KSA). The latter especially was designed to address the recent threat of the MERS-Cov virus. The developed model allows the

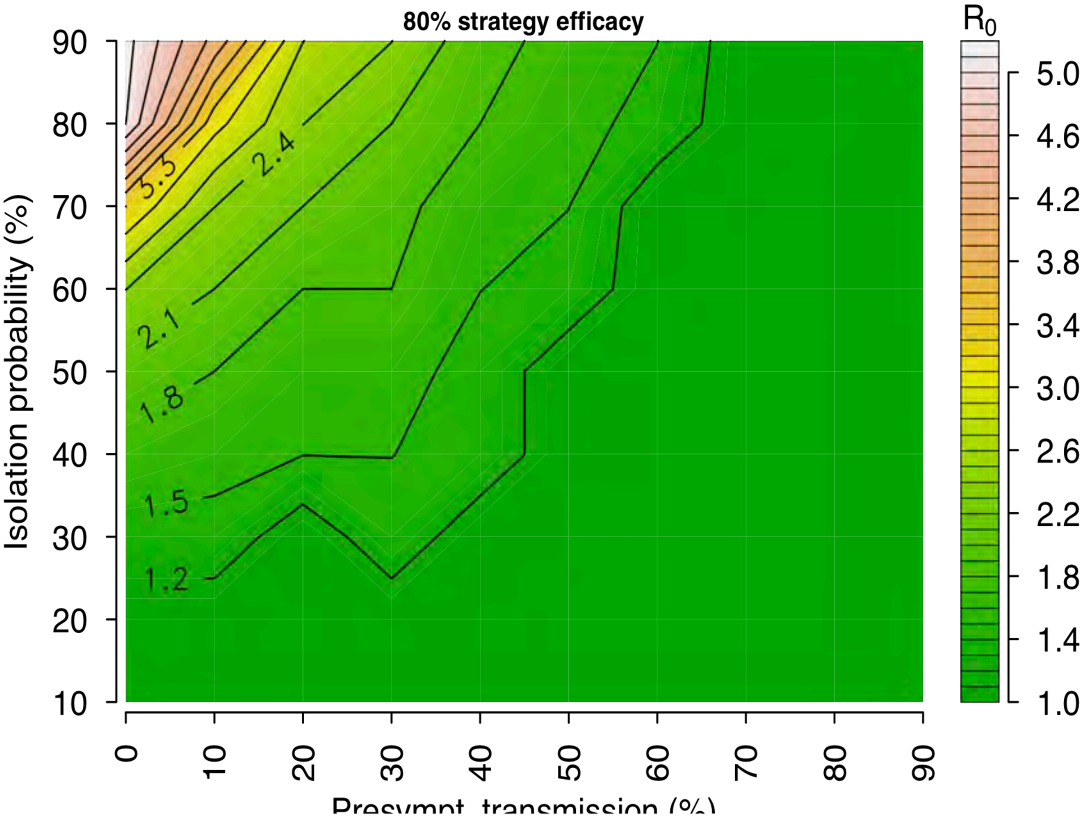


Fig 46: Contour plot showing the maximum R0 contained with 80% success in the case of SARS-like outbreaks as a function of the fraction of pre-symptomatic transmission and the case isolation rate.

analysis of the progression of the epidemic at the level of single individual and allows us to assess the likelihood of containment and mitigation as a function of the disease dynamic for a wide range of interventions including case isolation, household quarantine, schools and workplaces closure. A number of additional factors determine the effectiveness of containment policies, including how successful and timely is case isolation, the number of cases before implementation of control measures etc. For this reason, we performed a very extensive sensitivity analysis on many of the parameters that may affect the outbreak controllability. We focused on a parameter space for the dynamic of the disease that includes two basic regimes corresponding to influenza-like viruses and SARS-like viruses, respectively. Microsimulations results indicate that the case isolation rate is by far the most important parameter in controlling outbreaks while all others social distancing measures are generally producing incremental effects. Among the social distancing strategies, the most effective one is the reactive school closure. Moreover, social distancing effectiveness is depending on the socio-demographic setting. In Fig. 46 we show the maximum reproductive rate that can be contained as a function of the pre-symptomatic transmission and the effectiveness of the case isolation for the case of Mexico for SARS-like outbreak. It is possible to clearly see that effective containment for R0 >2.0 can be achieved only for low fraction of pre-symptomatic transmission. More in general, we provide a thorough analysis of outbreak controllability in a large region of parameters characterizing the dynamic of respiratory emerging infectious diseases. The controllability of outbreaks is hardly achievable for influenza-like pathogens even in the occurrence of combined isolation and social distancing strategies because of the high fraction of non-detectable transmission. For SARS-like viruses, the generation time and the proportion of asymptomatic transmission are in region of the parameter space where timely implemented non-pharmaceutical strategies may achieve containment probability of 90% or more. The results obtained here can be potentially used to discuss best control practices in for new emerging viruses such as the MERS-Cov.

We have also used the developed model to assess the risk that the accidental escape from a laboratory of a novel transmissible viral strain would not be contained in the local community [44]. We developed a model that specifically considers laboratory workers and their contacts in microsimulations of the epidemic onset See Fig. 47. We consider the following non-pharmaceutical interventions: isolation of the laboratory, laboratory workers' household quarantine, contact tracing of cases and subsequent household quarantine of identified secondary cases, and school and workplace closure both preventive and reactive. The Model simulations suggest that there is a non-negligible probability (5% to 15%), strongly dependent on reproduction number and probability of developing clinical symptoms, that the escape event is not detected at all. We find that the containment depends on the timely implementation of non-pharmaceutical interventions and contact tracing and it may be effective (>90% probability per event) only for pathogens with moderate transmissibility (reproductive number no larger than R0 = 1.5). Containment depends on population density and structure as well, with a probability of giving rise to a global event that is three to five times lower in rural areas. Results

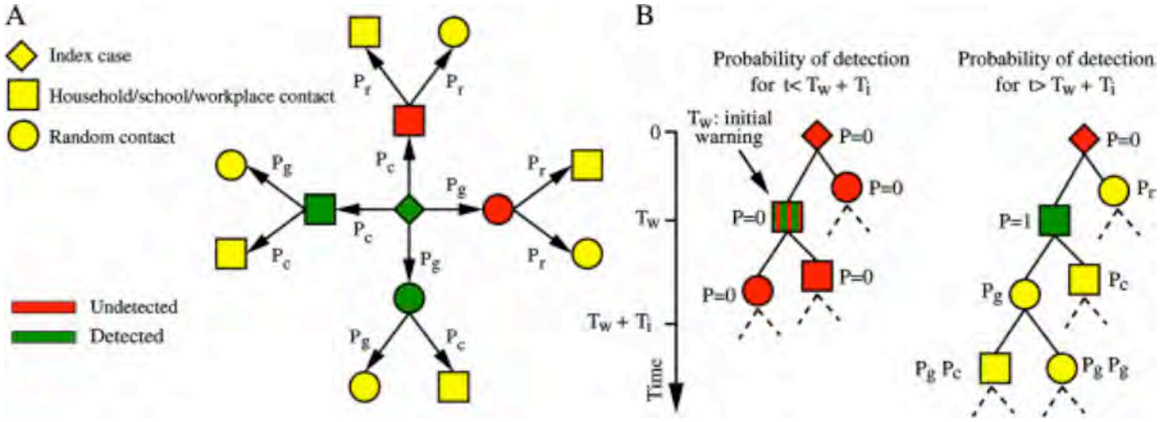


Fig 47: Contact tracing. (A) Probabilities of detecting first and second generation cases (the latter conditioned to the detection of first generation cases) triggered by a traced index case. (B) Example of network of cases triggered by the initial infected laboratory worker (undetected in this example; the initial warning is triggered by a secondary case in the laboratory), and probability of case detection at time of intervention (Tw + Ti). Tw is the time corresponding to the first identification of one of the initial cases. Ti is the timere quired to link the initial infections to an accidental release of the modified influenza strain in the laboratory and to activate the containment interventions.

suggest that controllability of escape events is not guaranteed and, given the rapid increase of biosafety laboratories worldwide, this poses a serious threat to human health. Our findings may be relevant to policy makers when designing adequate preparedness plans and may have important implications for determining the location of new biosafety laboratories worldwide.

4) Mathematical and statistical laws characterizing the non-local propagation of damage

In the context of remote assessment of damages in large-scale complex networks, we did start to investigate analytically and computationally the effectiveness in assessing the integrity, damage and residual function of large-scale networks by quick, targeted packet-probing of the network. We consider the traceroute tool, actually used in Internet probing as sampling algorithm, and we focus on different attack strategies, namely random or connectivity based.

Let us focus our attention in a network G(N,M), where N is the number of nodes and M is the set of vertices. We fix a number of sources $S = (s_1, s_2, \cdots, s_S)$ and a set of targets $T = (t_1, \cdots, t_T)$ among the N nodes. For each pair of target-source nodes the shortest path between joining the nodes is collected. After this operation is done for all source-target pairs, all the paths are merged resulting in sampled network G*(N*, M*), where the * represents the discovered portion of the network. We have developed a general approach based on a mean-field statistical analysis, which approximates the probability of edges and vertices to be detected as function of the number of sources, targets and the topological properties of the networks. In particular, we look at the case of damaged networks to develop the corresponding analytical and computational results quantifying the effectiveness in detecting the extent of network damage by packet probe sampling techniques. We can imagine the network damage as resulting from different attack strategies. The attacks can be random or targeted, according to some centrality measure. We indicate these two topological strategies as Dr and Dc where r stands for random and c for centrality measure. Another particular case of interest are WMD attacks that affect confined geographical areas. We indicate this local type of attack as Dg, where g stands for geographical. In order to asses the damage detection, a novel local quantity (per node) has been defined:

Mi=1-(NiD/Ni)

where Ni and NiD are the number of times that node i is detected in shortest path sampling the undamaged and damaged network respectively. If Mi=1 the node is not visible in the damaged network and very likely has been damaged. If Mi =0 the node is not damaged. If Mi<1 it implies that some of the shortest path present in the undamaged network have been compromised. These nodes are often at the periphery of the damaged region. Finally, Mi>1 implies that in the damaged network some nodes have become more central and thus are observed in more shortest paths than in the undamaged network. Those nodes are also potential candidates for overload and eventual avalanche of failures. The advantage of working with a local quantity is in the possibility of drawing geographically localized map as shown in Fig. 48.

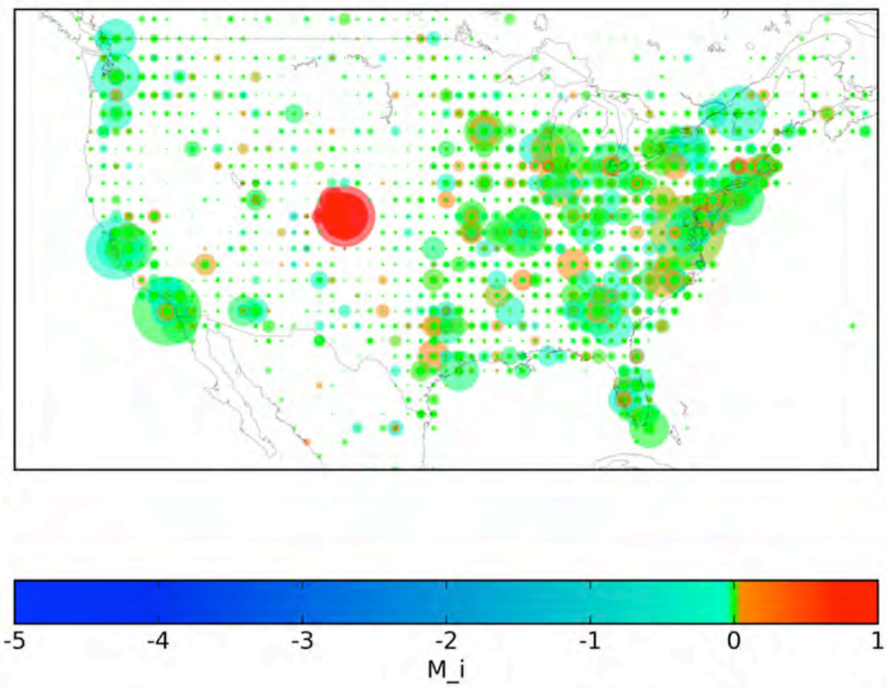


Fig 48: We show the damage assessed by a traceroute-like network exploration in the case of a large localized damage of the Internet autonomous system network in the area of Denver. It is possible to see the non-local effects (orange nodes) that ripple from the damaged area up to very distant nodes.

We can also introduce a global measure, M=ND/N (where ND is the number of nodes sampled in the damaged network and N is the number of nodes sampled in the un-damaged network) that allows to quickly identify damages that induce large variations in the routing patterns of networks. We then propose a statistical method able to classify un-sampled nodes as damaged or functioning. In this approach, we first sample a given input network through traceroute probes obtaining an approximated view of it. We then execute a controlled damage by removing nodes according to different strategies and sample the damaged network supposing not to know where the damage is. During the sampling, we constantly monitor whether the probability not to have seen a node exceeds the expected value calculated on the basis of the sampling of the undamaged graph. If this happens, we conclude that the node has been damaged, and henceforth will not be discovered in the following of the sampling. We test the performance of the method by counting how many nodes, among the real damaged ones, are detected as missing and how many are wrongly considered damaged even though they are not.

In the computational study we first focus on synthetic networks, either heterogeneous generated with the uncorrelated configuration model or homogeneous graphs obtained through the Erdos-Renyi model (ER). Remarkably, the entity and location of the damage can be detected with a good approximation. The accuracy improves for nodes that play an important role in network's connectivity. Namely, the detection of damaged hubs is more reliable than the one of peripheral nodes. As a practical application, we consider the damage detection on the physical Internet at the level of Autonomous System (AS). We use free sampled AS topology provided by the Dimes project In this case, to simulate real damages, we remove nodes according to their geographical displacement. In doing so we artificially reproduce critical events such as large-scale power outages or localized WMD attacks. Also in this case our method correctly identifies the location and the extent of the damage with great accuracy.

More precisely, the single node damage analysis is done by defining a statistical criterion that depends on the rate of observation given a rate of traceroute probing. We start again by monitoring the network {G} assuming that it is not damaged. The information gathered during the exploration process constitutes the null hypothesis of our measure, namely that no one of the nodes is damaged. Every time we send a traceroute probe we obtain a better approximation of the sampled network {G}* with increasing number of discovered nodes N*. At the same time we collect information about how many times a probe passes through a node i resulting on visit probability pi . The network is then damaged according to different strategies and sampled with traceroutes again. During the exploration a certain number of nodes are visited. By definition, these nodes are not damaged. The reason why a node is not seen can be either that it is actually damaged or that the sampling has missed it. To figure out which one of the two cases occurs we use p-value test applied to the visit probability p_i. Namely, we calculate the probability of not seeing the node i after a number T of traceroute probes sent as (1- p_i)T. The p-value test consist in imposing the equality between this quantity and an arbitrary confidence level C = (1- pi)T_i. By taking the logarithm on both side of the equation we obtain:

$$T_i = \ln C/\ln(1- p_i).$$

If the node i has not been seen at least once before Ti probes have been sent then we can state that i is damaged. Here we are assuming that the visit probability of nodes does not change too much after the damage. This holds when the damage is a relatively small perturbation and does not change the connectivity of the network or its dynamical properties.

The value of C tunes the number of probes to be sent before declaring a node damaged. If C is selected large, nodes will be declared as damaged early, leading to an overestimation of functioning nodes evaluated as damaged. We refer to these nodes as false positive damaged nodes (FPs). Conversely, if C is set to be small, more probes are needed to state if a node is damaged or not. The accuracy improves and the final response will eventually return only actually damaged nodes, the true positive damaged nodes (TPs). The optimal value of C is the one leading to the fastest but precise damage detection. We are finalizing a thorough set of simulations to identify optimal values of C in different basic synthetic network topologies that can be used a s a guidance in real networks. We also considered real Internet topology sample at the level of ASs where each node is an autonomous system of known geographical location and links represent the physical connections among them. Topologies are available for downloads in the DIMES project webpage in which each node is provided along with country and city it belongs to and with its geographical coordinates. We focus on the largest connected component of ASs that is constituted by 32,852 nodes. We test our method against two relevant classes of realistic attacks: all nodes in the same country are attacked, and all nodes inside a radius R with epicenter E are attacked. Both of these strategies are geography-based but they picture different scenarios. The first represents a deliberate shut down, for instance as happened in several countries during the Arab spring. The second one is referred to localized event such as WMD attack affecting an area of radius R.

We fix the number of probing sources to Ns = 15 and the density of sampled address at $r_T = 0.2$. According to one of the two strategies we remove ND nodes from the original AS network. The measure of damage detection should then be able to return not only the entity of the damage as a whole, but also tell us where the damage is localized. In order to achieve this goal we used the method discussed above. The choice of the confidence level is critical: a small confidence level corresponds to a very accurate measure of damaged nodes, while a big confidence level will produce a fast response.

As an example of entire country switch off we decided to damage all the nodes in Italy. Fig 49 shows the outcome of our method. We want to stress that the algorithm does not have any a priori information about where de damaged nodes are located. Despite that, the method clearly returns Italy as damaged country. Few other nodes are wrongly evaluated as damaged ones although they are not in Italy, and they constitute the FPs. One reason for the presence of FPs can be just statistical fluctuations. Another, more interesting, reason is that FP nodes turn out to be strongly linked to the Italian TPs, so that the deletion of the latter prevents them to be visited.

For the second type of geographical damage we decide to switch off all of ASs within a radius of 50 km around the city of Boston. Again the method is able to detect the damage in the right location as shown in Fig 50.

The choice for the optimal value of C is the one explained before, namely the larger value so that the precision is bigger than 0.9. In the case of damaging Italy this criterion sets C= 10-2 while for Boston C= 10-4. This difference is the result of the different damage extension. When damaging nodes within a certain radius, the visit of nodes outside that radius that are strongly related to the damaged ones is heavily affected, if not even prevented. This causes an overestimate of FPs that needs to be compensated by smaller value of C. Indeed, in this case it is interesting to note that FP nodes are located around the TP ones as a result of the inhibition effect induced by the damage.

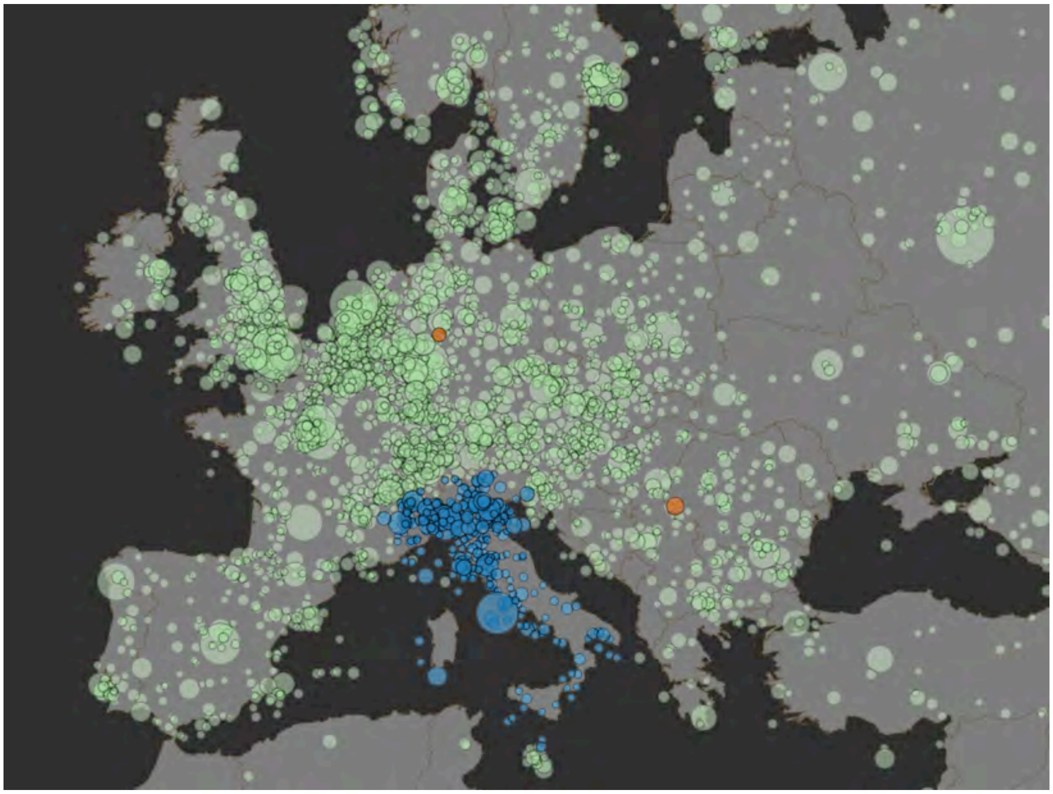


Fig 49: Analysis of the damage from recovered from the probing algorithm in the case of shut down of the entire Italian ASs network. Green circles are the working nodes, blue ones are the TPs while orange ones are the FPs.

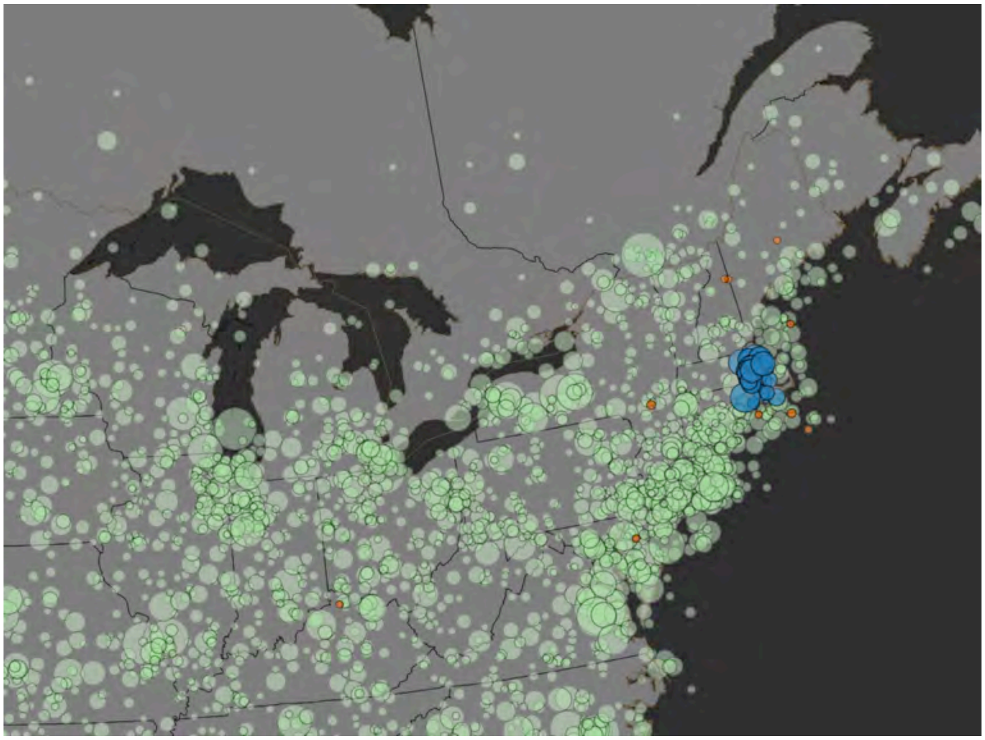


Fig 50: Map of part of the United States east coast showing the outcome after damaging nodes around the city of Boston within a radius of 50 km in the real AS network. Each circle represents one AS and its size is proportional to the degree. Green circles are the working ASs, blue ones are the TPs while orange ones are the FPs.

We believe that our method is able to identify the entity of the damage and more importantly its location. The method we have proposed can represent a first step towards a strategy for the continuous monitoring of large scale Infrastructure networks aimed to the quick detection of damages by remote locations. Importantly, we studied only random displacement of sources and targets for the damage discovery. Detection of damages could be improved by opportune choice of sources and targets or by different schedule of probes delivery. The results of this work have been published in Physical Review E [45].

5) Applications to real-world epidemic threats

While the work was focused primarily on WMD events, the theoretical understanding of socio-technical networks properties and dynamics and the modeling tools developed during the project are of more general interest and applications as proven by work done during the project for two major biological threats occurred during the life of the project: the H1N1-2009 pandemic, and the Ebola virus West Africa epidemic. It is also work remarking that the applications to real-world epidemic threats do not represent a deviation from the planned work, and have contributed to specific tasks of the project and produced advances to the planned work.

At the early stage in the life of the project we have been involved in the effort of the international scientific community to chart the unfolding of the 2009 H1N1 pandemic. We had the occasion to push side to side the research on multiscale reaction-diffusion networks while implementing and testing the results on the real time analysis of the H1N1 pandemic. The work done for the realistic modeling of the H1N1 pandemic has to be considered as a breakthrough that has shown for the first time, in a real world situation, the potential of computational methods in providing anticipations and forecasts that can be used in the support of the policy making and public health decision-making processes. The results obtained a very good agreement between predictions and real data, providing a strong and remarkable test of the quantitative level of the prediction offered by computational methods. The work carried out in developing computational techniques and simulations in multiscale networks were put to work to project the time course of the H1N1 flu epidemic in the United States. The simulations yielded projections and risk assessments of the epidemic outbreak based on the current knowledge of the disease parameters and took into account the human mobility networks on spatial scales between a few and a few thousand km.

The necessity to provide new way to obtain real-time estimates for the H1N1 parameters have pushed our team to work on a new methodology that perform a likelihood analysis of the model with respect to chronological data of the diffusion processes. Our methodology allowed early

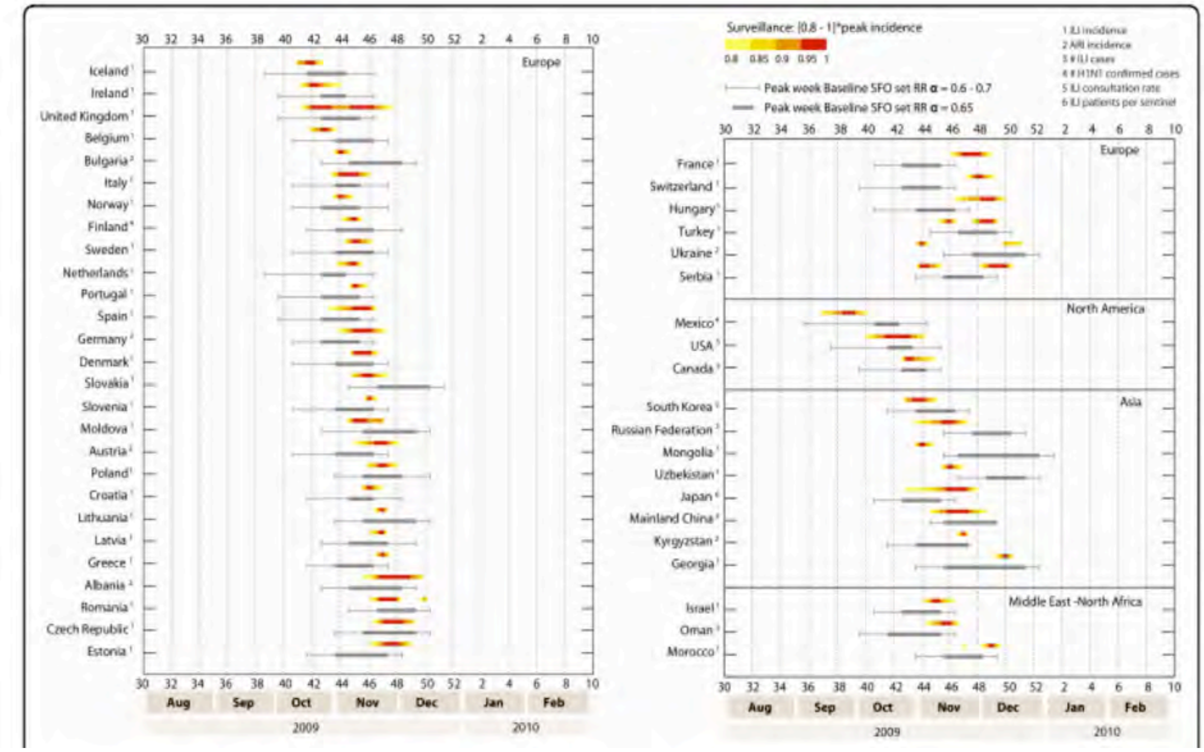


Fig 51: Peak timing in the northern hemisphere: simulations and real data. Peak weeks of the epidemic activity in the baseline stochastic forecast output (SFO) (gray). The reference ranges of the simulated peak week were obtained by analysis of 2,000 stochastic realizations of the model for three different values of the seasonal rescaling factor, α_{min}, of 0.6, 0.65, and 0.7. The peak weeks reported by the surveillance for the fall/winter wave are shown as color gradients, whose limits correspond to the time interval at which an incidence of greater than 80% of the maximum incidence was observed. The numbers 1 to 5 indicate the type of data provided by the surveillance of each country, and the numbered weeks of the year correspond to the calendar used by the US Center for Disease Control and Prevention.

estimates of the transmission potential of the H1N1 virus by taking advantage of the multi-scale diffusion processes defined by the population mobility networks. The developed model is coupling countries worldwide and this feature is extremely relevant in evaluating the time pattern of emerging infectious diseases. For instance, given a set of initial conditions for a local outbreak of new strain of influenza, the timeline of the arrival of the epidemic in each country and the ensuing activity peak is mainly determined by the human mobility network that couples different region of the world, as also discussed in our work on mitigation and containment of "secondary" Epidemic (E-WMD) threats. has the human mobility built-in, allowing to consistently simulate the mobility of infectious individuals on the global scale, thus providing ab-initio estimates of the epidemic timeline in each country or urban area without assumption on mobility and case importation. The multiscale model was calibrated with a maximum likelihood analysis fitting the data of the chronology of the H1N1 epidemic. This was done by generating several millions simulations on the worldwide scale of the pandemic evolution and finding the set of parameters that best fit the actual evolution of the pandemic. This allowed us to estimate the transmission potential of the disease and the seasonality features. The model has been used to provide predictions of the unfolding of the epidemic in the USA and Europe. The method anticipated an early peak occurring in October/November in most of the countries. The predictions, of a quantitative nature, have been published in early September 2009 on BMC Medicine [46]. The agreement between the predictions and the actual unfolding of the pandemic has been proven to be remarkable (see Fig 51), and have been validated against the real data provided by agencies of more than 40 countries in a paper published in BMC Medicine in 2012 [47]. The modeling effort has been used also for the analysis of vaccination campaigns, the effect of antiviral systematic use, the backtrack estimates of the actual number of cases in the Mexico, the projections for ICU occupancy and antibiotic usage [46], [48], [49].

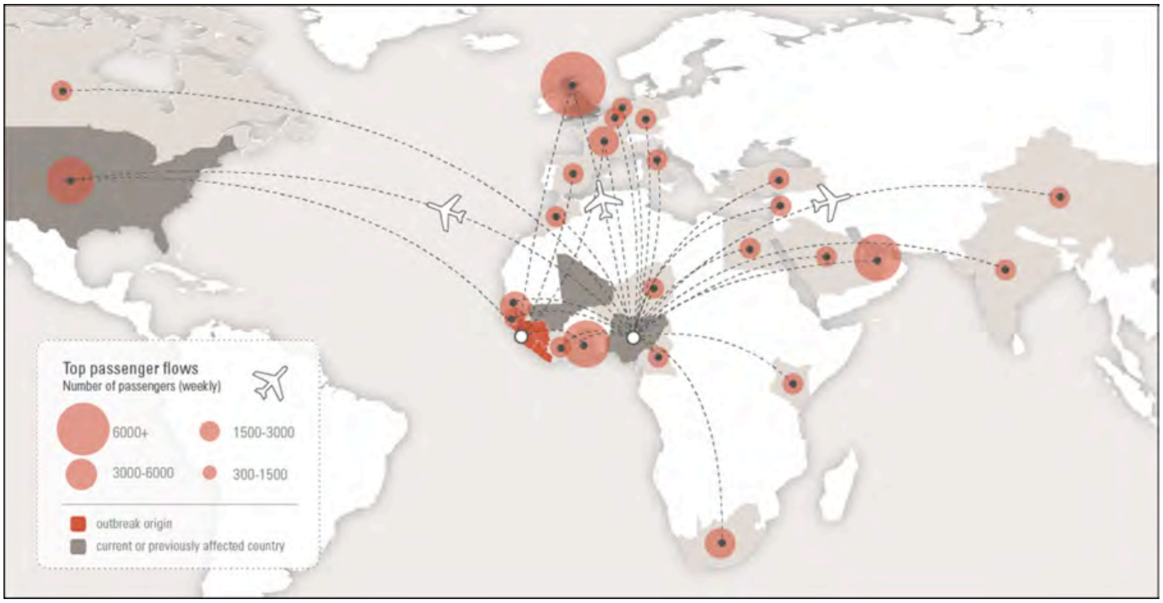
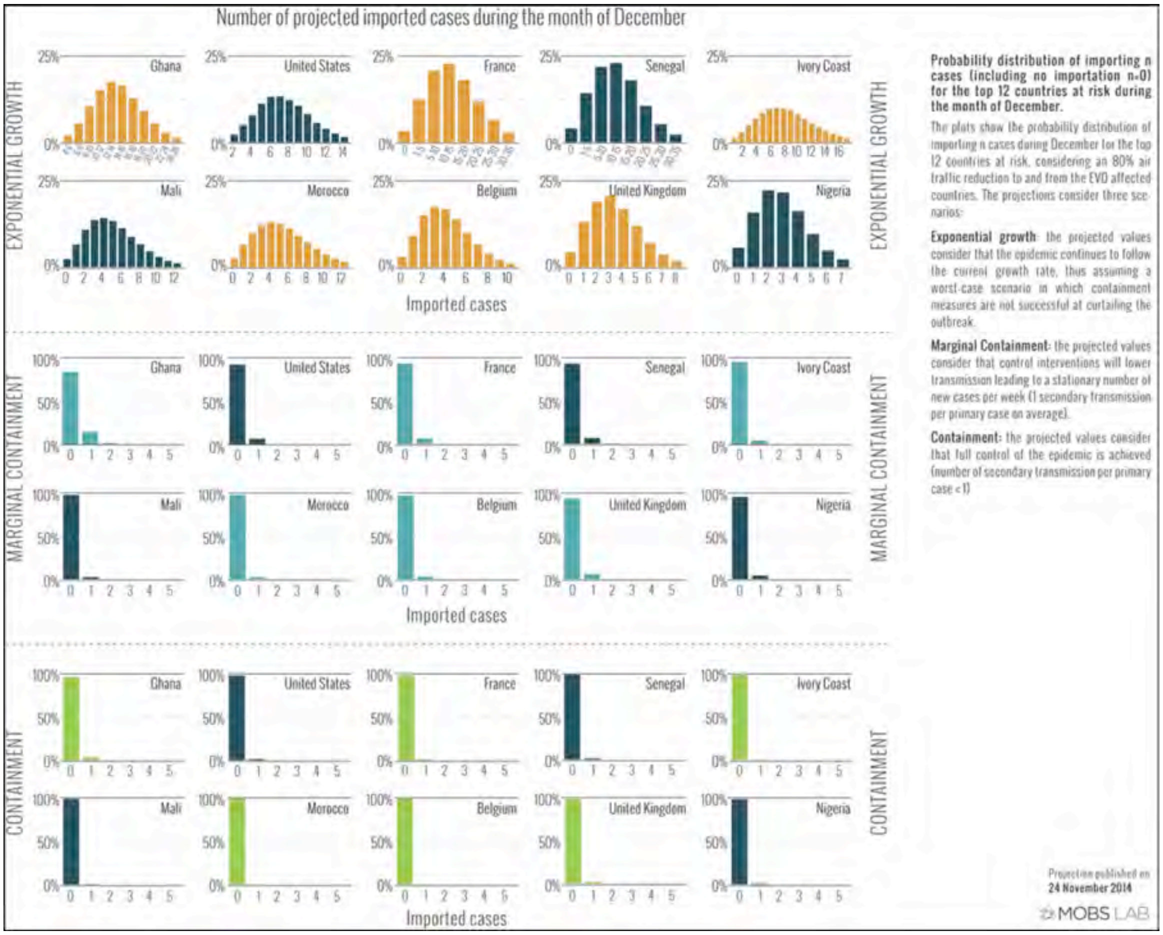

Top passenger flows
Number of passengers (weekly)
6000+
1500-3000
3000-6000
300-1500
outbreak origin
current or previously affected country

Fig 52: Air traffic connections from West African countries to the rest of the world

Another application of the work carried out during the life of the project is the use during the West Africa 2014 Ebola Virus Disease (EVD) epidemic of the multiscale network model and the network representation of origin-destination (O-D) data that explicitly track how many people travel from an origin location to a destination location regardless of the connections along the route (see Fig 52).

The 2014 West African Ebola Outbreak is so far the largest and deadliest recorded in history.

The most affected countries, Sierra Leone, Guinea, Liberia have been struggling to contain and to mitigate the outbreak. In July/August 2014 the rapid evolving situation of the Ebola Outbreak raised the need of real time analysis of the epidemic growth as well as an assessment of the risk of international dissemination, especially because the epidemic was affecting cities with major commercial airports. We have therefore used the results and tools developed in the past years of this grant to generate a 3-month ensemble forecast that provides quantitative estimates of the



Probability distribution of importing n cases (including no importation n=0) for the top 12 countries at risk during the month of December.

The plots show the probability distribution of importing n cases during December for the top 12 countries at risk, considering an 80% air traffic reduction to and from the EVD affected countries. The projections consider three scenarios:

Exponential growth: the projected values consider that the epidemic continues to follow the current growth rate, thus assuming a worst-case scenario in which containment measures are not successful at curtailing the outbreak.

Marginal Containment: the projected values consider that control interventions will lower transmission leading to a stationary number of new cases per week (1 secondary transmission per primary case on average).

Containment: the projected values consider that full control of the epidemic is achieved (number of secondary transmission per primary case < 1)

Fig 53 : Example of the projections routinely provided for the risk of international dissemination of the Ebola outbreak.

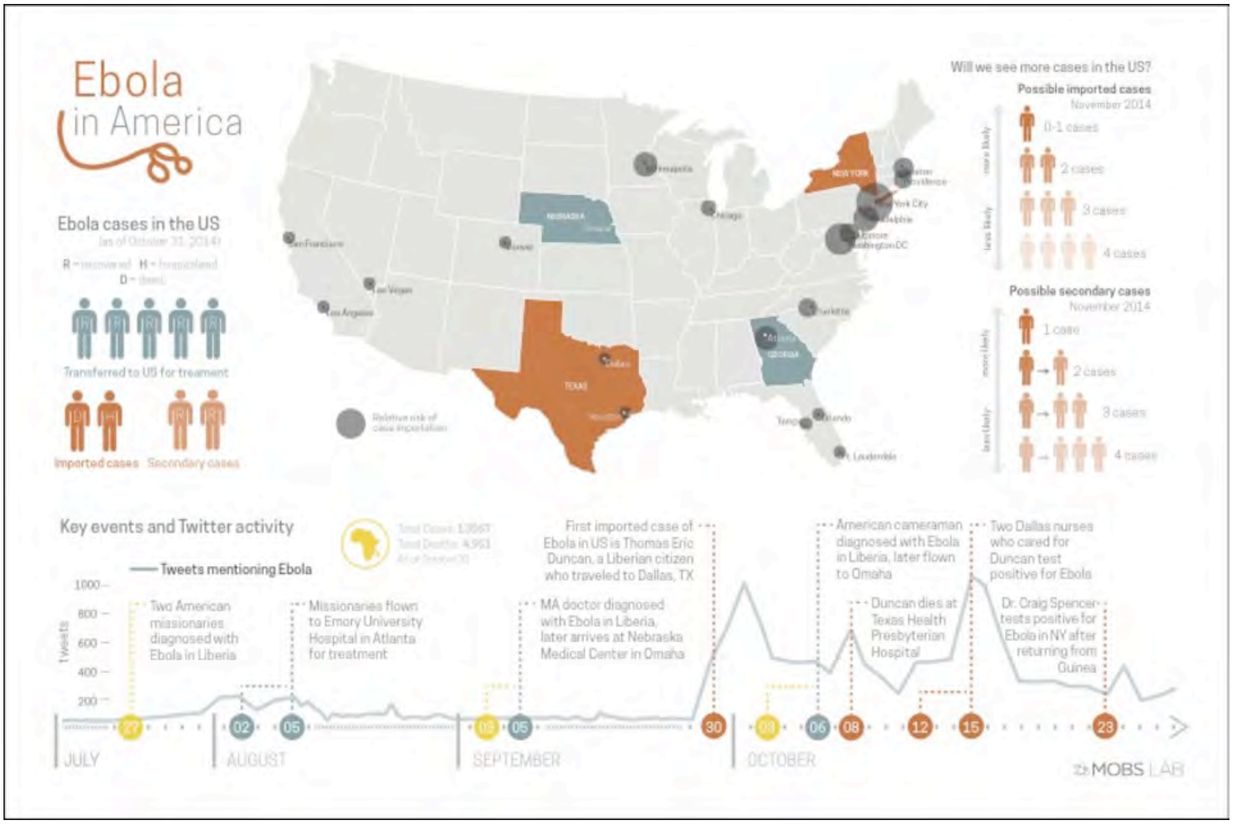


Fig 54 : Infographic summarizing some of the key results concerning the importation risk of ebola cases in the US.

local transmission of Ebola virus disease in West Africa and the probability of international spread if the containment measures are not successful at curtailing the outbreak. We simulated the international spreading of the outbreak and provide the estimate for the probability of Ebola virus disease case importation in countries across the world. In assessing case importation by airline travel it is also important to consider if the EVD cases are detected during a connecting flight or at the final destination of the traveler. In our analysis we have implemented two different models: i) the EVD cases are identified after the first connecting flight; ii) the EVD cases are able to travel to their final destination. These models provide a minimum and maximum for the probability of case importation in each country, whose spread depends on whether the country's transportation systems act as a traffic gateway or a destination hub. This modeling application has allowed us working with "segments" and O-D data in the case of airline mobility. We have initiated a thorough comparison of the worldwide segment and O-D airline transportation networks, assessing the impact of different data sources and the risk assessment process.

The need for continuous updates of the results, as more and more data became available from the affected region, led to the development of a web repository (http://www.mobs-lab.org/ebola.html) where we have been continuously updating our initial results. The modeling

activity focused on providing the probability of observing imported ebola cases in the US and across the world, the expected number of cases and their probability distribution (see Figs 53 and 54). Specific work has been carried out also for the analysis of travel restrictions. Our results have been published in PLOS Currents Outbreaks [50] and Eurosurveillance [47], both journals devoted to the rapid dissemination of results concerning epidemic threats. The results contained in these papers and the ensuing updates disseminated via web have been prominently featured in the media and discussed during an oversight hearing in the U.S. House of Representatives, Thursday, October 16, 2014.
(http://news.sciencemag.org/health/2014/10/congressional-ebola-debate-invokes-plos-paper)

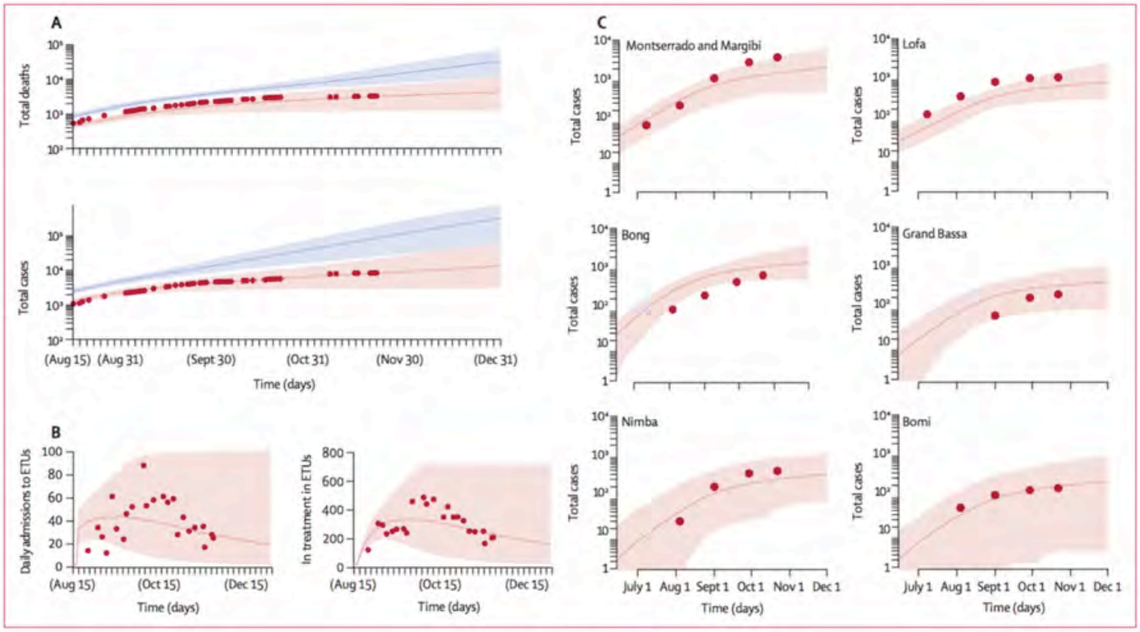


Fig 55: Spatio-temporal dynamic of the Ebola epidemic in Liberia. Data and modeling results comparison.

In order to improve the spatial resolution and realism of the within-country analysis we have capitalized network models developed in the past years of this grant to develop a spatial agent-based model that integrates socio-demographic data for Liberia to estimate the relative importance of the main settings for Ebola virus disease transmission. The model is based on synthetic population in which every household in Liberia is explicitly represented and includes hospitals, clinics and Ebola treatment units and the risk of spread to health-care workers working at them. We used the model to project the spatiotemporal spreading of the disease and to disentangle the effect of Ebola treatment unit availability, safe burial procedures, and the distribution of household protection kits in the aversion of Ebola virus disease cases (see Fig 55). This work has been published in Lancet Infectious Diseases [51], providing a modeling platform able to characterize the spatiotemporal spreading pattern of the Ebola outbreak down to the level of specific counties. The agent-based model has been refined and extended to Guinea and Sierra Leone for additional analysis and the assessment of the effect of ring vaccination trials in the elimination of the disease. The application of our work to the Ebola outbreak analysis has brought to the attention of the community the possibility of using large-scale multiscale network and agent based models for real-time forecast and situational awareness, and represents one of the success stories of the work carried out during the life of this project.

References

[1] F. Simini, M. C. González, A. Maritan, and A.-L. Barabási, "A universal model for mobility and migration patterns.," *Nature*, pp. 8–12, 2012.

[2] Y. Ren, M. Ercsey-Ravasz, P. Wang, M. C. González, and Z. Toroczkai, "Predicting commuter flows in spatial networks using a radiation model based on temporal ranges.," *Nat. Commun.*, vol. 5, p. 5347, 2014.

[3] D. Balcan, V. Colizza, B. Gonçalves, H. Hu, J. J. Ramasco, and A. Vespignani, "Multiscale mobility networks and the spatial spreading of infectious diseases.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 51, pp. 21484–21489, 2009.

[4] A. Vespignani, "Predicting the behavior of techno-social systems," *Science (80-.).*, vol. 325, no. July, pp. 425–428, 2009.

[5] A. Vespignani, "Modelling dynamical processes in complex socio-technical systems," *Nat. Phys.*, vol. 8, no. 1, pp. 32–39, 2012.

[6] M. Ercsey-Ravasz and Z. Toroczkai, "Centrality scaling in large networks," *Phys. Rev. Lett.*, vol. 105, no. 3, pp. 2–5, 2010.

[7] M. Ercsey-Ravasz, R. N. Lichtenwalter, N. V. Chawla, and Z. Toroczkai, "Range-limited centrality measures in complex networks," *Phys. Rev. E*, vol. 85, no. 6, p. 066103, 2012.

[8] L. Witten, TA; Sanders, "Diffusion-limited aggregation, a kinetic critical phenomenon," *Phys. Rev. Lett.*, vol. 47, no. 19, 1981.

[9] F. Barra, B. Davidovitch, A. Levermann, and I. Procaccia, "Laplacian Growth and Diffusion Limited Aggregation: Different Universality Classes," *Phys. Rev. Lett.*, vol. 87, no. 13, p. 134501, 2001.

[10] D. Bauer, H. Broersma, and E. Schmeichel, "Toughness in graphs - A survey," *Graphs Comb.*, vol. 22, pp. 1–35, 2006.

[11] M. Ercsey-Ravasz, Z. Toroczkai, Z. Lakner, and J. Baranyi, "Complexity of the international agro-food trade network and its impact on food safety.," *PLoS One*, vol. 7, no. 5, p. e37810, 2012.

[12] Wikipedia, "List of cognitive biases - Wikipedia, the free encyclopedia:" [Online]. Available: https://en.wikipedia.org/wiki/List_of_cognitive_biases. [Accessed: 08-Dec-2015].

[13] E. T. Jaynes, "Information theory and statistical mechanics," *Phys. Rev.*, vol. 106, no. 4, pp. 620–630, 1957.

[14] E. T. Jaynes, "Information theory and statistical mechanics II," *Phys. Rev.*, vol. 108, no. 2, pp. 171–190, 1957.

[15] S. Pressé, K. Ghosh, J. Lee, and K. a. Dill, "Principles of maximum entropy and maximum caliber in statistical physics," *Rev. Mod. Phys.*, vol. 85, no. 3, pp. 1115–1141, 2013.

[16] D. Strauss, "On a general class of models for interaction," *SIAM Rev.*, vol. 28, no. 4, pp. 513–527, 1986.

[17] S. Horvát, É. Czabarka, and Z. Toroczkai, "Reducing Degeneracy in Maximum Entropy Models of Networks," *Phys. Rev. Lett.*, vol. 114, no. 15, p. 158701, 2015.

[18] H. Kim, Z. Toroczkai, P. L. Erdős, I. Miklós, and L. Á. Székely, “Degree-based graph construction,” *J. Phys. A Math. Theor.*, vol. 42, no. 39, p. 12, 2009.

[19] C. I. Del Genio, H. Kim, Z. Toroczkai, and K. E. Bassler, “Efficient and exact sampling of simple graphs with given arbitrary degree sequence,” *PLoS One*, vol. 5, no. 4, pp. 1–8, 2010.

[20] P. L. Erdős, I. Miklós, and Z. Toroczkai, “A simple Havel-Hakimi type algorithm to realize graphical degree sequences of directed graphs,” *Electron. J. Comb.*, vol. 17, p. #R66, 2009.

[21] H. Kim, C. I. Del Genio, K. E. Bassler, and Z. Toroczkai, “Constructing and sampling directed graphs with given degree sequences,” *New J. Phys.*, vol. 14, no. 2, p. 023012, 2012.

[22] P. Erdös, I. Miklós, and Z. Toroczkai, “A Decomposition Based Proof for Fast Mixing of a Markov Chain over Balanced Realizations of a Joint Degree Matrix,” *SIAM J. Discret. Math.*, vol. 29, no. 1, pp. 481–499, Jan. 2015.

[23] K. E. Bassler and T. Z. Genio, Charo I Del, Erdős Péter L, Miklós István, “Exact sampling of graphs with prescribed degree correlations,” *New J. Phys.*, vol. 17, no. 8, p. 083052, 2015.

[24] C. Orsini, M. M. Dankulov, P. Colomer-de-Simón, A. Jamakovic, P. Mahadevan, A. Vahdat, K. E. Bassler, Z. Toroczkai, M. Boguñá, G. Caldarelli, S. Fortunato, and D. Krioukov, “Quantifying randomness in real networks,” *Nat. Commun.*, vol. 6, no. May, p. 8627, 2015.

[25] A. Asztalos and Z. Toroczkai, “Network Discovery by Generalized Random Walks,” *Europhys. Lett.*, vol. 92, no. December, p. 50008, 2010.

[26] C. WANG, O. LIZARDO, D. HACHEN, A. STRATHMAN, Z. TOROCZKAI, and N. V. CHAWLA, “A dyadic reciprocity index for repeated interaction networks,” *Netw. Sci.*, vol. 1, no. 01, pp. 31–48, 2013.

[27] M. Ercsey-Ravasz, N. T. Markov, C. Lamy, D. C. Van Essen, K. Knoblauch, Z. Toroczkai, and H. Kennedy, “A Predictive Network Model of Cerebral Cortical Connectivity Based on a Distance Rule,” *Neuron*, vol. 80, no. 1, pp. 184–197, 2013.

[28] N. T. Markov, M. M. Ercsey-Ravasz, A. R. Ribeiro Gomes, C. Lamy, L. Magrou, J. Vezoli, P. Misery, A. Falchier, R. Quilodran, M. A. Gariel, J. Sallet, R. Gamanut, C. Huissoud, S. Clavagnier, P. Giroud, D. Sappey-Marinier, P. Barone, C. Dehay, Z. Toroczkai, K. Knoblauch, D. C. Van Essen, and H. Kennedy, “A Weighted and Directed Interareal Connectivity Matrix for Macaque Cerebral Cortex,” *Cereb. Cortex*, vol. 24, no. 1, pp. 17–36, 2014.

[29] N. T. Markov, P. Misery, a. Falchier, C. Lamy, J. Vezoli, R. Quilodran, M. a. Gariel, P. Giroud, M. Ercsey-Ravasz, L. J. Pilaz, C. Huissoud, P. Barone, C. Dehay, Z. Toroczkai, D. C. Van Essen, H. Kennedy, and K. Knoblauch, “Weight consistency specifies regularities of macaque cortical networks,” *Cereb. Cortex*, vol. 21, no. 6, pp. 1254–1272, 2011.

[30] N. T. Markov, M. Ercsey-Ravasz, C. Lamy, A. R. Ribeiro Gomes, L. Magrou, P. Misery, P. Giroud, P. Barone, C. Dehay, Z. Toroczkai, K. Knoblauch, D. C. Van Essen, and H. Kennedy, “The role of long-range connections on the specificity of the macaque interareal cortical network.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, no. 13, pp. 5187–92, 2013.

[31] H. Kennedy, K. Knoblauch, and Z. Toroczkai, “Why data coherence and quality is critical for understanding interareal cortical networks,” *Neuroimage*, vol. 80, pp. 37–45, 2013.

[32] W. Van den Broeck, C. Gioannini, B. Gonçalves, M. Quaggiotto, V. Colizza, and A. Vespignani, “The GLEaMviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale.,” *BMC Infect. Dis.*, vol. 11, no. 1, p. 37, 2011.

[33] M. Ajelli, B. Gonçalves, D. Balcan, V. Colizza, H. Hu, J. J. Ramasco, S. Merler, and A. Vespignani, “Comparing large-scale computational approaches to epidemic modeling: agent-based versus structured metapopulation models.,” *BMC Infect. Dis.*, vol. 10, p. 190, 2010.

[34] A. Pastore Y Piontti, M. F. D. C. Gomes, N. Samay, N. Perra, and A. Vespignani, “The infection tree of global epidemics,” *Netw. Sci.*, vol. 2, no. 01, pp. 132–137, 2014.

[35] D. Balcan and A. Vespignani, “Invasion threshold in structured populations with recurrent mobility patterns,” *J. Theor. Biol.*, vol. 293, pp. 87–100, 2012.

[36] D. Balcan and A. Vespignani, “Phase transitions in contagion processes mediated by recurrent mobility patterns,” *Nat. Phys.*, vol. 7, no. 7, pp. 581–586, 2011.

[37] D. Balcan, B. Gonçalves, H. Hu, J. J. Ramasco, V. Colizza, and A. Vespignani, “Modeling the spatial spread of infectious diseases: The GLobal Epidemic and Mobility computational model,” *J. Comput. Sci.*, vol. 1, no. 3, pp. 132–145, 2010.

[38] N. Perra, D. Balcan, B. Gonçalves, and A. Vespignani, “Towards a characterization of behavior-disease models.,” *PLoS One*, vol. 6, no. 8, p. e23084, 2011.

[39] S. Meloni, N. Perra, A. Arenas, S. Gómez, Y. Moreno, and A. Vespignani, “Modeling human mobility responses to the large-scale spreading of infectious diseases,” *Sci. Rep.*, vol. 1, pp. 1–7, art no 00062, 2011.

[40] C. Poletto, S. Meloni, A. Van Metre, V. Colizza, Y. Moreno, and A. Vespignani, “Characterising two-pathogen competition in spatially structured environments,” *Sci. Rep.*, vol. 5, p. 7895, 2015.

[41] C. Poletto, S. Meloni, V. Colizza, Y. Moreno, and A. Vespignani, “Host Mobility Drives Pathogen Competition in Spatially Structured Populations,” *PLoS Comput Biol*, vol. 9, no. 8, p. e1003169, 2013.

[42] B. Gonçalves, D. Balcan, and A. Vespignani, “Human mobility and the worldwide impact of intentional localized highly pathogenic virus release.,” *Sci. Rep.*, vol. 3, p. 810, 2013.

[43] P. Bajardi, C. Poletto, J. J. Ramasco, M. Tizzoni, V. Colizza, and A. Vespignani, “Human mobility networks, travel restrictions, and the global spread of 2009 H1N1 pandemic.,” *PLoS One*, vol. 6, no. 1, p. e16591, 2011.

[44] S. Merler, M. Ajelli, L. Fumanelli, and A. Vespignani, “Containing the accidental laboratory escape of potential pandemic influenza viruses.,” *BMC Med.*, vol. 11, p. 252, 2013.

[45] F. Ciulla, N. Perra, A. Baronchelli, and A. Vespignani, “Damage detection via shortest-path network sampling,” *Phys. Rev. E*, vol. 89, no. 5, p. 052816, 2014.

[46] D. Balcan, H. Hu, B. Goncalves, P. Bajardi, C. Poletto, J. J. Ramasco, D. Paolotti, N. Perra, M. Tizzoni, W. Broeck, V. Colizza, and A. Vespignani, “Seasonal transmission potential and activity peaks of the new influenza A(H1N1): a Monte Carlo likelihood analysis based on human mobility,” *BMC Med.*, vol. 7, no. 1, p. 45, 2009.

[47] C. Poletto, M. Gomes, A. Pastore y Piontti, L. Rossi, L. Bioglio, D. Chao, I. Longini, M. Halloran, V. Colizza, and A. Vespignani, “Assessing the impact of travel restrictions on international spread of the 2014 West African Ebola epidemic,” *Eurosurveillance*, vol. 19, no. 42, p. 20936, 2014.

[48] R. J. Colizza V, Vespignani A, Perra N, Poletto C, Gonçalves B, Hu H, Balcan D, Paolotti D, Van den Broeck W, Tizzoni M, Bajardi P, “Estimate of ovel Influenza A / H1 1 cases in Mexico at the early stage of the pandemic with a spatially structured epidemic model,” *PLOS Curr. Influ.*, pp. 1–9, 2010.

[49] D. Balcan, V. Colizza, A. C. Singer, C. Chouaid, H. Hu, B. Gonçalves, P. Bajardi, C. Poletto, J. J. Ramasco, N. Perra, M. Tizzoni, D. Paolotti, W. Van den Broeck, A.-J. Valleron, and A. Vespignani, “Modeling the critical care demand and antibiotics resources needed during the Fall 2009 wave of influenza A(H1N1) pandemic.,” *PLoS Curr.*, vol. 1, p. RRN1133, 2009.

[50] M. F. C. Gomes, A. Pastore y Piotti, L. Rossi, D. Chao, I. Longini, and M. E. Halloran, “Assessing the International Spreading Risk Associated with the 2014 West African Ebola Outbreak,” *PLOS Curr. Outbreaks*, pp. 1–22, 2014.

[51] S. Merler, M. Ajelli, L. Fumanelli, M. F. C. Gomes, A. P. Y. Piontti, L. Rossi, D. L. Chao, I. M. J. Longini, M. E. Halloran, and A. Vespignani, “Spatiotemporal spread of the 2014 outbreak of Ebola virus disease in Liberia and the effectiveness of non-pharmaceutical interventions: a computational modelling analysis.,” *Lancet. Infect. Dis.*, vol. 3099, no. 14, pp. 1–8, 2015.

DEPARTMENT OF DEFENSE

DEFENSE THREAT REDUCTION
AGENCY
8725 JOHN J. KINGMAN ROAD
STOP 6201
FORT BELVOIR, VA 22060
ATTN: P. TANDY

DEFENSE TECHNICAL
INFORMATION CENTER
8725 JOHN J. KINGMAN ROAD,
SUITE 0944
FT. BELVOIR, VA 22060-6201
ATTN: DTIC/OCA

DEPARTMENT OF DEFENSE CONTRACTORS

QUANTERION SOLUTIONS, INC.
1680 TEXAS STREET, SE
KIRTLAND AFB, NM 87117-5669
ATTN: DTRIAC